RICE UNIVERSITY

Essays on Treatment Effects Evaluation

by

Ronghua Guo

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Robin C. Sickles, Chair
Reginald Henry Hargrove Chair of
Economics,
Rice University

Natalia Sizova
Assistant Professor of Economics,
Rice University

David W. Scott
Noah Harding Professor of Statistics,
Rice University

HOUSTON, TEXAS
April 2012

RICE UNIVERSITY

Essays on Treatment Effects Evaluation

by

User

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Robin C. Sickles, Chair
Reginald Henry Hargrove Chair of Economics,
Rice University

Natalia Sizova
Assistant Professor of Economics,
Rice University

David W. Scott
Noah Harding Professor of Statistics,
Rice University

HOUSTON, TEXAS
April 2012

ABSTRACT

Essays on Treatment Effects Evaluation

Ronghua Guo

The first chapter uses the propensity score matching method to measure the average impact of insurance on health service utilization in terms of office-based physician visits, total number of reported visits to hospital outpatient departments, and emergency room visits. Four matching algorithms are employed to match propensity scores. The results show that insurance significantly increases office-based physician visits, and its impacts on reported visits to hospital outpatient departments and emergency room visits are positive, but not significant. This implies that physician offices will receive a substantial increase in demand if universal insurance is imposed. Government will need to allocate more resources to physician offices relative to outpatient or emergency room services in the case of universal insurance in order to accommodate the increased demand.

The second chapter studies the sensitivity of propensity score matching methods to different estimation methods. Traditionally, parametric models, such as logit and probit, are used to estimate propensity score. Current technology allows us to use computationally intensive methods, either semiparametric or nonparametric, to estimate it. We use the Monte Carlo experimental method to investigate the sensitivity of the

treatment effect to different propensity score estimation models under the unconfoundedness assumption. The results show that the average treatment effect on the treated (ATT) estimates are insensitive to the estimation methods when index function for treatment is linear, but logit and probit model do better jobs when the index function is nonlinear.

The third chapter proposes a Cross-Sectionally Varying (CVC) Coefficient method to approximate individual treatment effects with nonexperimental data, the distribution of treatment effects, the average treatment effect on the treated and the average treatment effect. The CVC method reparameterizes the outcome of no treatment and the treatment effect in terms of observable variables, and uses these observables together with a Bayesian estimator of their coefficients to approximate individual treatment effects. Monte Carlo simulations demonstrate the efficacy and applicability of the proposed estimator. This method is applied to two datasets: data from the U.S. Job Training Partnership ACT (JTPA) program and a dataset that contains firms' seasoned equity offerings and operating performances.

Acknowledgments

First and foremost I want to express my deepest gratitude to my advisor Robin Sickles, who has supported me throughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way. I attribute the level of my Doctor degree to his encouragement and effort, he has been my inspiration as I hurdle all the obstacles in the completion of this research work. One could not wish for a better or friendlier supervisor. Sincere thanks to Mahmoud El-Gamal, Natalia Sizova and David Scott for serving on my committee and guiding me to the palace of econometrics and statistics. I truly appreciate their instruction and firm support throughout the years, which allows me overcome every challenge I met. I am also very grateful to P.A.V.B. Swamy for introducing the cross-sectionally varying coefficient model to me, and providing me substantial help on technical skills. I-lok Chang and Yamil Kaba, who help me on the programing, also deserve special thanks.

I want to thank Altha Rodger for her cheerful assistance over the years, she is the best coordinator. Of course, I would also like to say sincere thanks to all my fellow graduate students. Thanks for your accompany and sharing of your experience about life, research and career.

Last but not least, my parents have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice. Their love and encouragement are my endless source of energy. Tons of thanks go to my beloved husband, Xin. His mathematics knowledge, life care and spiritual support create

the best environment for me to study and do research. He fills my life with sunshine. It would have been much harder to finish my Ph. D without him.

For any errors or inadequacies that may remain in this work, the responsibility is entirely my own.

Contents

List of Figures

List of Tables

Chapter 1

The Impact of Health Insurance on Health Service Utilization

1.1. Introduction

Most people in the United States are not able to pay for medical care like doctor office visits out of their own pockets, due to the high costs of health care. So health insurance comes in to share patients' costs. Common sense suggests that individuals with health insurance coverage will be more likely to get access to provider's services than the uninsured, which in turn increases utilization, while the uninsured might not go to the hospitals until their health status deteriorates severely. Because the number of uninsured people is rising, scholars heatedly debate who the uninsured people are and ways that the uninsured could obtain health insurance coverage. Over the past few years, more than 15 percent of the population in United States has no health insurance coverage; this lack of health insurance coverage has been blamed for the large disparity in access to health care services. Many Americans, who lack health insurance, forgo health care, especially

preventive care. Forgoing health care often results in more costly stages outcomes. Thus, the delay of health care not only increases national health costs, but also leads to long-term health declines, shorter life expectancy, and even bankruptcies due to the inability to pay medical bills. While the influences of other factors (i.e., the social determinants of health) are compelling, it is clear that universal access to health insurance will markedly reduce health disparities.

Buchmueller,Grumbach, Kronick and Hahn (2005) noted that "Health care reform aimed at establishing universal insurance coverage has been a topic of policy debate on and off for the past half century. Following the demise of the Clinton administration's health care reform initiative, discussions of comprehensive reforms were replaced with a focus on more narrowly defined policies, such as the expansion of public coverage for children and enhanced regulation of private insurance markets." On March 23, 2010, President Barack Obama signed Patient Protection and Affordable Care Act (PPACA) and started comprehensive reform of health insurance. PPACA, taking effect in 2010, is designed to eliminate pre-condition screening and premium loading, and removes lifetime and annual coverage caps. It also creates price competition to lower insurance rate, provides more subsidies to the poor to make insurance affordable to them, and puts in place an individual mandate which is formally called the shared responsibility requirement and which legislates that everyone must have health insurance. All of those provisions aim to expand nationwide health insurance coverage.

Increasing insurance coverage would of course usher in challenges to health care providers. If resources not allocated to the right divisions, one division might encounter difficulty in taking all patients with its limited capacity even while another division has

idle resources. Understanding the probable magnitude of insurance's impacts on the health care utilization is very important for evaluating the costs and benefits of expansion strategies. It is also crucial for resources allocation.

Common sense suggests that having health insurance would make people use the health services more, as they will need to pay significantly less than if they were uninsured. Previous researchers have studied the impact of health insurance coverage among the indigent, children, and the elderly. LaPlante (1993) uses data from the 1989 National Health Interview Survey to evaluate the health insurance coverage of children and nonelderly adults with disabilities and their utilization of physician and hospital care as a function of health insurance status. He found that uninsured adults with disabilities have from 19 percent to 44 percent fewer physician contacts than similar adults with insurance. Currie and Gruber (1996) explore expansions of the Medicaid program to low-income children with the National Health Interview Survey and study the effect of public insurance for children on their utilization of medical care and health outcomes. Currie and Gruber find that the Medicaid expansions increased visits in a relatively efficient way: most of the visits took place at the doctor's office rather than in emergency rooms. French and Kamboj (2002) also insist that health care utilization depend upon access to health insurance. Card, Dobkin, and Maestas (2008) estimate the effect of insurance coverage on health care utilization and health outcomes by studying the onset of eligibility for the Medicare program at age 65. They find that there is a large increase in self-reported access to health care and doctors' visits at age 65 for less educated minority people who typically did not have insurance coverage before 65, which suggests that insurance coverage does affect their health care utilization. Kolstad and Kowalski (2010)

investigate the impacts of mandatory insurance on insurance coverage, hospital utilization patterns, preventive care, quality of care and hospital cost growth with data for the state of Massachusetts, which passed legislation aimed at achieving near universal health insurance coverage in April 2006, and other control states. They show that the number of inpatient admissions originating from the emergency room decreased, with some evidence also suggesting an increase in the utilization of preventive services and a decline in hospitalizations for preventable conditions.

In this chapter, we analyze three types of health services: self-reported office-based physician visits, total number of reported visits to hospital outpatient departments, and emergency room (ER) visits using the Medical Expenditure Panel Survey (MEPS) data for nonelderly adults (of age between 18 and 64). The impacts of insurance on these three variables show the disparities between insured and uninsured people and indicate which service faces significant changes under universally mandated insurance. Office-based physicians are one of the crucial components of the community healthcare system, fundamentally assuring the health of the local community. The number of doctor visits is also a key measure of access to health care; both private plans and public programs have coverage for it. Use of the emergency room (ER) can be thought as a way of entry for inpatient care. Because hospitals must provide at least some care, without regard to insurance status, the emergency room is a potentially major access to hospital care for the uninsured. The emergency room is designed to treat acute and urgent health events. If it is a patient's primary point of care, then the patient might not access to preventive care that could mitigate future severe health events. Uninsured individuals who get inpatient care after a visit to the emergency room also have barriers to receiving follow up

treatment, which would reduce the efficacy of the inpatient care they received from the emergency room. Knowing the differences in these three variables between the insured and uninsured groups is crucial for policy makers.

Since the uninsured could differ from the insured people in both observable and unobservable characteristics, it is difficult to draw causal inferences from these types of simple comparisons. Furthermore, insurance coverage itself could be a function of health status, leading to an endogeneity problem in estimating of the impacts of insurance on health care utilization. In this paper, we use propensity score matching to estimate the treatment effect of health insurance on medical care use, which overcome the selection bias problem. The chapter is structured as follows: Section 2 describes the data and outlines the statistical summary for both insured and uninsured groups. We illustrate the treatment effect model and how the classical unconfoundedness assumption is employed to solve selection bias for the model in section 3. The propensity score matching method is provided in section 4, in which we describe the common ways to estimate propensity score, four kinds of matching algorithms, and balance test of the propensity score. Section 5 analyzes the health care utilization problem with propensity score matching method provided that the insurance is the treatment; the results show insurance has significant positive impact on office-based physician visits, but not for the other two measures of utilization. Section 7 concludes this chapter.

1.2. Data

1.2.1. Data Description

We use data from The Medical Expenditure Panel Survey (MEPS), a nationally representative sample of the non-institutionalized population. This survey, starting in 1996, is a set of large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.), and employers across the United States. It collects detailed data on health services spending and health insurance; furthermore, it links that information to the demographic characteristics, employment, income, health status, and other characteristics of survey respondents. Additionally, MEPS provides information on insurance status at several different times that facilitate the identification of changes in health insurance status during a year. Here, we use only the 1996 Full Year Consolidated Data File to look at the treatment effect of insurance on health care service utilization.

As children's utilization pattern is highly likely to differ systematically from that of adults, and elderly adults (age over 64) receive universal insurance coverage through Medicare program, we will focus on the non-elderly adults (ages 18 to 64) in this paper. The status of health insurance coverage is dynamic. It could change once one individual lost his job and employer-sponsored health insurance. It could also change if some insured people decided to change their current plan and had no insurance coverage temporarily. Rising cost could force the low-income group to quit their insurance plan. In trying to estimate the treatment effect of insurance on health service utilization, we focus on individuals whose insurance status was stable over the year 1996 (either fully covered

by insurance or no insurance over the year); this approach ignores people who change their insurance status over the year.

Table 1.1 Sample Mean and Standard Errors of Covariates for Insured and Uninsured Groups

Variable	Insured (Treated)	Uninsured (Control)
fams	2.923 (1.478)	3.392 (1.684)
region	2.576 (1.031)	2.627 (1.047)
MSA	0.808 (0.394)	0.770 (0.421)
age	40.990 (10.216)	37.673 (12.122)
gender	0.441 (0.497)	0.534 (0.499)
race	4.836 (2.335)	4.835 (2.372)
marry	0.660 (0.474)	0.639 (0.481)
educ	13.596 (2.492)	12.583 (2.713)
wage	33342.95 (22001.67)	19705.41 (18409.41)
health	2.021 (0.920)	2.118 (0.970)
OFVST	3.071 (4.936)	2.426 (5.180)
OUTVST	0.4442 (0.4442)	0.3203 (0.3203)
ERVST	0.1318 (0.4498)	0.1665 (0.5288)
Sample Size	4370	2841

To control other characteristics of each observation, we have taken into account family size, geographical region, age, gender, race, education year, wage, perceived health status, and whether they are from rural or urban area (MSA in Table 1.1). Health care service utilization is measured by self-reported office-based physician visits (OFVST in Table 1.1), total number of reported visits to hospital outpatient departments (OUTVST in Table 1.1), and emergency room visits (ERVST in Table 1.1).

1.2.2. Distributions of Insured and Uninsured Samples

Table 1.1 provides the characteristics of the sample we use, with 4370 insured observations and 2841 uninsured observations. Family size (*fams*) gives the number of people in the respondent's family. *Region* takes the value of 1 if respondent is in the northeast United States, 2 if in the midwest, 3 if in the south, 4 if in the west. *MSA* takes the value of 1 if respondent is from a metropolitan area and 0 otherwise. If the respondent is female, *gender* is set to be 1, for male, it is set as 0. Variable *Race* has five different values: 1 for American Indian, 2 for Aleut or Eskimo, 3 for Asian or Pacific Islander, 4 for Black and 5 for White. Marital status (*marry*) is 1 for married, and 0 otherwise. Yearly wage is measured in dollars. Health status (*health*) has value from 1 to 5, which mean excellent, very good, good, fair and poor respectively. *Educ*, with min=1 and max=17, describes the education year for each respondent. *OFVST* represents office-based physician visits in year 1996, and *OUTVST* is visits to hospital outpatient department, while *ERVST* is for emergency room visits. From the table, insured people tend to have more years of education, higher wages and more office-based physician visits. Another interesting point is that insured people tend to have fewer emergency room visits than the uninsured ones. To the extent that lack of insurance led people to use

the emergency room as a point of entry for treatment that they otherwise would have sought through another channel, we would expect to see a smaller number of inpatient admissions originating in the emergency room for the insured individuals. It could imply that uninsured people might postpone going to hospital until the last minute, when they are very ill. However, the differences of the health care utilizations between insured and uninsured group cannot be the pure result brought by insurance, because those two groups differ in average education, wage etc., which affect their use of health service and are also influential factors in the insurance decisions. The ultimate goal of this paper is to check if holding insurance will have any significant impact on health service usage (office-based physician visit etc.) by adjusting for the health care visits that stem from other factors besides insurance status. In this case, we will solve this problem with treatment effect model.

1.3. Treatment Effect Model

The treatment effect model is used to estimate the effect of a binary treatment, T, where T=1 for an active treatment and T=0 for a control treatment. Let Y_1 denote a response variable that would be observed under the active treatment and Y_0 denote a response variable under the control treatment. The two variables are called potential outcomes (Neyman 1923; Rubin, 1974, 1977). If we denote by X a vector of pre-treatment variables, each unit under study is sampled from the joint distribution of (Y, X, T), where$Y = TY_1 + (1 - T)\, Y_0$. For the analysis with MEPS data, Y would be the office-based physician visits (or visits to hospital outpatient department, emergency room

visits), treatment T represents whether or not the individual has health insurance, and all other characteristics as age, gender and etc. are the pre-treatment variables X.

The literature pay a lot of attention on evaluation of the average treatment effect on the treated (ATT), which is also called causal effect. Assume that a random sample of N units, indexed i=1, …, N, is drawn from a large population, causal effect is defined as

$$\gamma_{ATT} = E(Y_{i1} - Y_{i0} \,|T_i = 1) = \; E(Y_{i1} \,|T_i = 1) - E(Y_{i0} \,|T_i = 1) \qquad (1.3.1)$$

Since for each unit under study Y_{i1} and Y_{i0} are never jointly observed, we face what is called the fundamental problem of causal inference (Holland 1986, Holland and Rubin 1988). The fundamental problem of causal inference implies that such a unit level causal effect can never be calculated with the observed data. In order to estimate ATT, a proper substitute has to be chosen for the counterfactual result Y_{i0}. Using the mean outcome of untreated units $E(Y_{i0} \,|T_i = 0)$ is usually not a good idea, because it is most likely that factors, which determine the treatment decision, also affect the outcome variable of interest. Thus, Y_i and T_i are correlated, leading to a self-selection bias shown as below. For ATT, it can be noted as:

$$E(Y_{i1} \,|T_i = 1) - E(Y_{i0} \,|T_i = 0) = \gamma_{ATT} + \; E(Y_{i0} \,|T_i = 1) - E(Y_{i0} \,|T_i = 0) \qquad (1.3.2)$$

The difference between the left hand side of equation (3.2) and γ_{ATT}

$$E(Y_{i0} \,|T_i = 1) - E(Y_{i0} \,|T_i = 0)$$

is the so called "self-selection bias." The true parameter γ_{ATT} is only identified if

$$E(Y_{i0} \,|T_i = 1) - E(Y_{i0} \,|T_i = 0) = 0 \qquad (1.3.3)$$

Another parameter of interest is the average treatment effect (ATE), which is defined as

$$\gamma_{ATE} = E(Y_{i1} - Y_{i0}) \tag{1.3.4}$$

The challenge when estimating ATE is that both counterfactual outcomes $E(Y_{i1}\,|T_i = 0)$ and $E(Y_{i0}\,|T_i = 1)$ have to be constructed. To assess ATT and ATE, we must reply on comparisons of the potential outcomes across units. In social experiments where assignment to treatment is random, (1.3.3) is ensured, and the treatment effect is identified. In non-experimental studies, one has to invoke some identifying assumptions to solve the problem stated in equation (1.3.1) and (1.3.4). The following assumption extends the framework to non-experimental settings.

Assumption 1a (Unconfoundedness**)** Rosenbaum and Rubin (1983)

For each unit *i* with pre-treatment covariates X_i, $(Y_{i1}, Y_{i0}) \perp T_i \mid X_i$.

This assumption is also referred as "independence assumption." It implies that selection is solely based on observable characteristics X_i, and that all variables that influence treatment assignment (decision) and potential outcomes simultaneously are observed by the researcher. This is a strong assumption but is fundamentally untestable.

1.1. Propensity Score Matching Method

By the unconfoundedness assumption, comparing two individuals with the same observable attributes, one of whom was treated and one of whom was not, is like comparing those two individuals in a randomized experiment. Then the treatment effect for the treated, γ_{ATT}, is identified: it is equal to the treatment effect conditional on

covariates and assignment (decision) to treatment, γ_{ATT} |($X_i, T_i = 1$), averaged over the distribution of $X_i \mid T_i = 1$.

One way to do the estimation would be matching units on their vector of covariates, X_i. In principle, we could stratify the data into sub-groups (or bins), each defined by a value of X_i; within each bin, take the difference of $Y_{i1}\ and\ Y_{i0}$. The limitation of this method is that it requires a sufficiently rich comparison group so that no bin contains a treated unit without a comparison unit. However, this is always impossible in reality. Rosenbaum and Rubin (1983) presented a useful extension that replace multivariate X_i with a scalar function of X_i, called the propensity score, which gives the probability of taking treatment at each value of X_i.

Assumption 1b (Unconfoundedness Based on Propensity Score)

Let p(X_i) be the probability of an individual i having been assigned to treatment (or having decided to take the treatment), defined as $\mathrm{p}(X_i) = \Pr(T_i = 1 \mid X_i) = E(T_i|X_i)$. Then

$$(Y_{i1}, Y_{i0}) \perp T_i \mid X_i \implies (Y_{i1}, Y_{i0}) \perp T_i \mid p(X_i) \tag{1.4.1}$$

Using the propensity score substantially reduces the dimensionality problem of matching on X_i, it allows us to match on a scalar variable p(X_i) rather than in a general n-space X_i.

Assumption 2 (Overlap)

$$0 < p(X_i) < 1 \tag{1.4.2}$$

It ensures that individuals with the same X_i have a positive probability of being participants or non-participants. Rosenbaun and Rubin (1983) call the combination of the two assumptions as "strong ignorability." With "strong ignorability", γ_{ATT} could be written as

$$\gamma_{ATT} = E_{\mathrm{p}(X_i)\,|T=1}\{E\big(Y_{i1}\,\big|T_i = 1, \mathrm{p}(X_i)\big) - E\big(Y_{i0}\,\big|T_i = 0, \mathrm{p}(X_i)\big) \tag{1.4.3}$$

The propensity score matching estimator of γ_{ATT} is simply the mean difference in outcomes of treated and control units over the common support, appropriately weighted by the propensity score distribution of observation; and treated and control observations are matched on the propensity scores.

1.1.1. Estimation of Propensity Score

To estimate the propensity score, we generally have to solve the following model, and get $\mathrm{p}(X_i) = \Pr(T_i = 1 \mid X_i)$ as the estimated propensity score.

$$T_i = 1 \text{ If } T_i^* > 0$$

$$T_i = 0 \text{ Otherwise}$$

$$T_i^* = g(X_i) + \epsilon_i$$

where g is unknown, and ϵ_i is independent of X_i.

Most of the work applying this have concentrated on logit (or probit) model for ease of estimation. Both logit and probit models assume g takes the linear functional form of X_i as $X_i'\beta$, but they hold different assumptions on ϵ.

The logit model assumes ϵ has a symmetric logistic distribution, in this case,

$$\mathrm{p}(X_i) = E(T_i \,|X_i) = \frac{1}{1 + e^{\lambda X_i' \beta}}$$

The probit model assumes $\epsilon \sim N(0, \sigma^2)$, in this case,

$$\mathrm{p}(X_i) = E(T_i \,|X_i) = \phi(\lambda X_i' \beta)$$

Here, T_i is the treatment status indicating if the individual has insurance coverage, and X_i are the covariates on which we want to match to obtain an ignorable treatment assignment.

1.1.2. Matching Methods

An estimate of the propensity score is only the first step in estimating treatment effect. The propensity score is a continuous variable between 0 and 1; theoretically we have a probability of zero to observe two units with exactly the same value of propensity score. We need to use matching algorithm to connect treated and control individuals.

Different matching algorithms have been proposed in the literature. Four of them are Nearest Neighbor Matching, Radius Matching, Kernel Matching and Local Linear Matching. These four algorithms are widely used in applications.

Nearest Neighbor Matching:

Let T denote the set of all treated units and C the set of all control units, and Y_{i1} and Y_{j0} be the observed outcomes of the treated and control units respectively. To make the notation neat, $\mathrm{p}(X_i)$ is referred as p_i in the following discussion. $C(i)$ stands for the

set of control units matched to the treated unit i with an estimated propensity score of p_i. Nearest neighbor matching set

$$C(i) = \min_j | p_i - p_j |$$

$C(i)$ is a singleton set unless there are multiple nearest neighbors. In practice, the case of multiple nearest neighbors should be very rare, in particular if the set of characteristics X_i contains continuous variables.

We denote the number of controls matched with observation $i \in T$ by N_{i0} and define the weights $w_{ij} = \frac{1}{N_{i0}}$ if $j \in C(i)$ and $w_{ij} = 0$ otherwise. Then, the formula for the ATT estimator is written as follows:

$$\gamma_{ATT} = \frac{1}{N_1} \sum_{i \in T} \{ Y_{i1} - \sum_{j \in C(i)} w_{ij} Y_{i0} \} \quad (1.4.4)$$

N_1 is the number of treated units.

Radius Matching

In radius matching,

$$C(i) = \{ p_j \mid | p_i - p_j | < r \}$$

All the control units with estimated propensity scores falling within a radius r from p_i are matched to the treated unit i. And the ATT estimator is

$$\gamma_{ATT} = \frac{1}{N_1} \sum_{i \in T} \{ Y_{i1} - \sum_{j \in C(i)} w_{ij} Y_{i0} \} \quad (1.4.5)$$

Kernel Matching

The kernel matching estimator is given by

$$\gamma_{ATT} = \frac{1}{N_1}\sum_{i\in T}\left\{Y_{i1} - \frac{\sum_{j\in C}\ G(\frac{p_j - p_i}{h_n})Y_{i0}}{\sum_{k\in C}\ G(\frac{p_k - p_i}{h_n})}\right\}$$

where $G(\cdot)$ is a kernel function and h_n is a bandwidth parameter. In terms of equation (1.4.4) and (1.4.5), the weighting function, w_{ij} is equal to $\frac{G(\frac{p_j - p_i}{h_n})}{\sum_{k\in C}\ G(\frac{p_k - p_i}{h_n})}$. Under standard conditions of the bandwidth and kernel, $\frac{\sum_{j\in C}\ G(\frac{p_j - p_i}{h_n})Y_{i0}}{\sum_{k\in C}\ G(\frac{p_k - p_i}{h_n})}$ is a consistent estimator of $E\big(Y_{i0}\ \big|T_i = 0, \mathrm{p}(X_i)\big)$.

Local Linear Matching

Heckman, Ichimura and Todd (1997) propose a generalized version of kernel matching, called local linear matching. The local linear weighting function is given by

$$w_{ij} = \frac{G_{ij}\sum_{k\in C}\ G_{ik}\ (p_k - p_i)^2 - [G_{ij}\big(p_{kj} - p_i\big)[\sum_{k\in C}\ G_{ik}\ (p_k - p_i)]}{\sum_{j\in C} G_{ij}\sum_{k\in C}\ G_{ik}\ (p_k - p_i)^2 - (\sum_{k\in C}\ G_{ik}\ (p_k - p_i))^2}$$

The Nearest Neighbor method suffers from poor matches, because, for some particular treated units, the nearest neighbor would find a matching unit with a far different propensity score, which would affect the estimation of the treatment effect. The Radius Matching and Kernel Matching methods overcome this problem. Radius Matching performs matching only if the control units have estimated propensity scores falling in a pre-determined neighborhood of the propensity score of the treated unit. The

smaller the size of the neighborhood the better is matching quality, as the size of radius is the upper bound of the distance between treated unit and the matched control units. If the radius is set very small, some treated units will not be matched because the neighborhood does not contain control units. With Kernel Matching, all treated are matched with a weighted average of all control units; the weights are inversely proportional to the distance between the propensity scores of treated and controls. Fan (1992a, b) demonstrated that local linear estimation has some advantage over standard kernel estimation; it has a faster rate of convergence near boundary points and greater robustness to different data design densities. Thus, local linear regression would be expected to perform better than kernel estimation in cases where the nonparticipant observations on p_i fall on one side of the participant observations.

Implementing the asymptotic standard error formulae for those matching estimators is cumbersome, so standard errors are often generated by bootstrap resampling methods. Abadie and Imbens (2006) show that standard bootstrap resampling methods are not valid for assessing the variability of the nearest neighbor matching estimators, although applicable to the kernel and local linear matching estimators. They give alternative standard error formulae for deriving the standard error of nearest neighbor method. We use R package "Matching" by Sekhon (2011) to calculate the nearest neighbor matching, as it incorporates the valid formula of standard deviation derived by Abadie and Imbens. For the rest of the methods, we use the standard error from bootstrap resampling.

Matching can be carried on with and without replacement. Performances depend on the data and in particular on the degree of overlap between the treatment and control

groups in terms of the propensity scores. When there is substantial overlap in the distribution of the propensity scores between the treatment and control groups, most of the matching algorithms will yield similar results. When the treatment and control units are remarkably different, finding a satisfactory match by matching without replacement can be very problematic. In particular, if there are only a handful of comparison units comparable to the treated units, then once these comparison units have been matched, the remaining treated units will have to be matched to comparison units that are far different. In such settings matching with replacement is the natural choice. If there are no comparison units for a range of propensity scores, then for that range the treatment effect cannot be estimated. All the results in the paper are based on matching with replacement.

1.1.3. Common Support and Balance Test

As matching has to be performed to satisfy the overlap assumption, it is vital to check the overlap of support between treatment and control group. Several ways are suggested in the literature. Most straightforward one is the visual observation of the distributions of the propensity scores in treated and control groups. Lechner (2000b) argues that there is no need to implement a complicated, formal estimator, given that the support problem can be spotted by inspecting the propensity score distribution. However, some formal methods have been developed. One of them compares the minima and maxima of propensity scores for the treatment and control groups. Smith and Todd (2005) suggest a trimming method to determine the common support. Implementing the common support condition ensures that any combination of characteristics observed in the treatment group can also be observed in the control group (Bryson, Dorsett, and

Purdon, 2002). For ATT, this is sufficient to ensure the existence of potential matches in the control group.

The primary purpose of the propensity score is that it serves as a balancing score. Thus, the idea behind balancing tests is to check whether the propensity score is an adequate balancing score. That is, to check if X have the same distribution for the treatment and comparison groups at each value of the propensity score. More formally, we are interested in verifying if $T \perp X \mid p(X)$. The intuition behind the notation is that additional knowledge of X should not provide new information on T, if we have already considered information brought by p(X). The propensity scores themselves only serve as tools to balance the observed distribution of covariates across the treated and comparison groups. Therefore, the success of propensity score estimation is assessed by the balancing test rather than by the fitness of the models used to estimate propensity scores.

1.2. Empirical Results

Region, MSA, gender, race, marry, health, fams, age, educ and *wage* constitute the pre-treatment (or observable) variables for this analysis of insurance's impact on the health service utilization. The propensity score matching method gives ATT results in table 1.2. Numbers, in parenthesis, are standard deviations. The ATT significant at 5% is marked with an asterisk. In order to make the estimated propensity score balanced, we include these terms: *region, MSA, gender, race, marry, health, fams, fams*^{2}*, age, age*^{2} *educ, wage, wage*^{2} and *wage *age*. The distributions of estimated propensity scores for treated and control group can be found in Figure 1.1, which shows propensity scores for the treated group occur rarely at the boundary of 0 and 1. Estimates with and without

common support for different matching methods are in Table 1.2. They do not differ much for each matching algorithm, owing to the rich range of counterparts in the control group. Among the four matching algorithms, radius, kernel and local linear method yield similar results, while the nearest neighbor method gives an estimate a little far from them. As we mentioned in section 1.4.2, the nearest neighbor matching method sometimes finds a matching unit with a far different propensity score, so it cannot estimate the ATT as well as other methods.

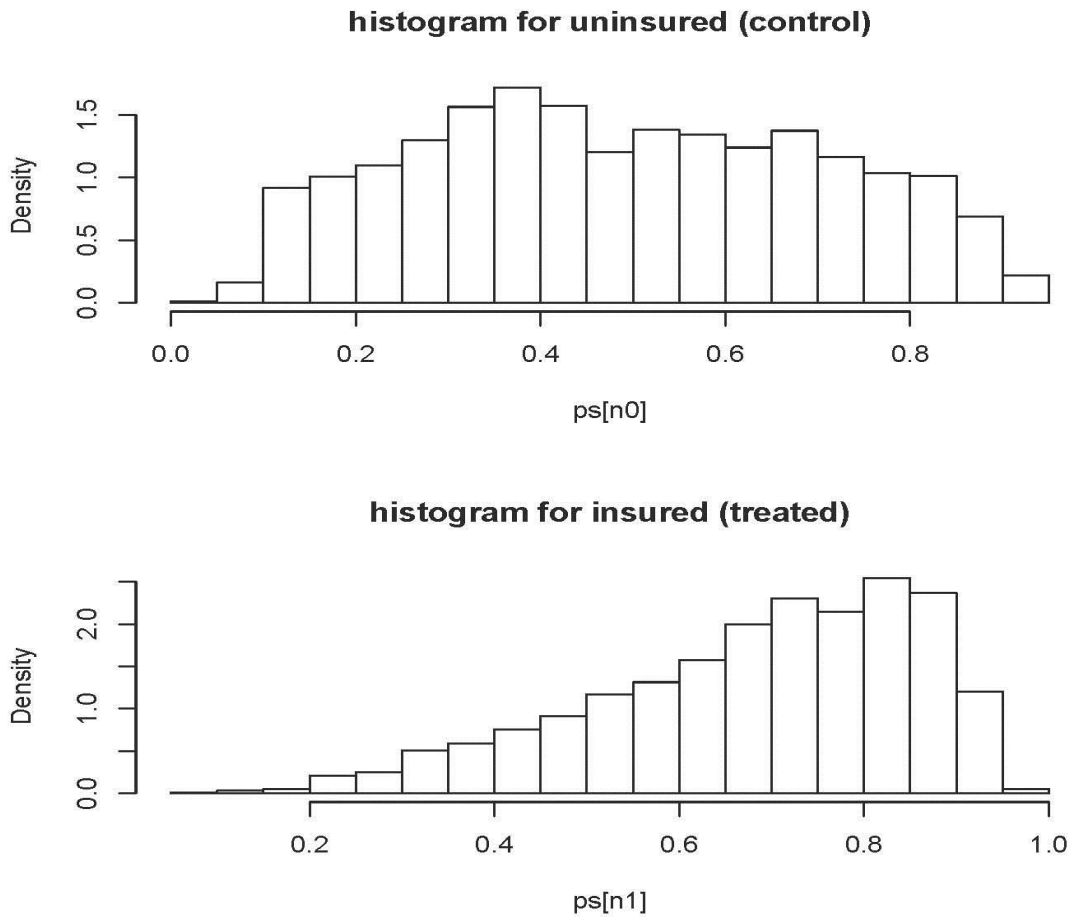

Figure 1.1 Histograms of propensity scores for insured and uninsured group

In the data description part, we noticed that insured group has higher mean values for office-based physician visits (OFVST) and hospital outpatient department visits (OUTVST), but lower mean for emergency room visit (ERVST). The latter is contrary to the common sense that insured people will use health services more, as they pay less (or no) money out of their pocket. Given that we control pre-treatment variables, will this trend disappear? Table 1.2 confirms common sense: the causal effect of insurance on emergency room visits is positive, which means insured people tend to go to the emergency room more. As this causal effect is not significant, we conclude that emergency room usage is almost the same for insured and uninsured people. So is hospital outpatient department usage. As for office-based physician visit, insurance has significant positive causal impact on it. The lack of insurances introduces disparity in office-based physician visits, but neither in the emergency room nor the outpatient department utilization.

Table 1.2 ATT of insurance on different variables

Matching Method	OFVST	OUTVST	ERVST
Nearest neighbor	0.5554* (0.2191)	0.1096 (0.0890)	0.0213 (0.0185)
Nearest neighbor with common support	0.5394* (0.2170)	0.1062 (0.0884)	0.0210 (0.0184)
Radius (0.01)	0.3554* (0.1576)	0.1235 (0.0748)	0.0018 (0.0157)
Radius (0.01) with common support	0.3484* (0.1569)	0.1241 (0.0747)	0.0013 (0.0156)
Kernel (h=0.01)	0.3585* (0.1564)	0.1200 (0.0743)	0.0027 (0.0156)
Kernel (h=0.01) with common support	0.3582* (0.1551)	0.1179 (0.0740)	0.0024 (0.0155)
Local linear (h=0.01)	0.3988* (0.1616)	0.1195 (0.0761)	0.0031 (0.0161)
Local linear (h=0.01) with common support	0.3562* (0.1560)	0.1191 (0.0743)	0.0021 (0.0156)

1.3. Conclusion

This chapter estimates the causal effect of insurance on health care utilization based on the propensity score matching method. The result shows insurance has a significant positive effect on office-base physician visits, which means uninsured people might not visit a doctor because of high medical cost. However, we do not have enough evidence to conclude that insurance makes any difference on outpatient department visits and emergency room visits.

Chapter 2

Sensitivity of Propensity Score Matching Method to Different Propensity Score Estimation Algorithms

2.1. Introduction

Estimating treatment effects suffers from the selection bias problem (see Heckman 1979). Matching, which is a method to match treated observations to comparisons with similar covariates, is becoming a popular procedure to correct the selection bias under the assumption of unconfoundedness. This assumption is also known as conditional independence. Because a bias is introduced by the difference of characteristics between the treated group and the comparison group, matching on covariates is a straightforward way to correct the bias. Rubin (1980) states that matching

on covariates by definition will remove the difference of characteristics and the bias. The matching method, compared with regression-type estimators, would release some of assumptions for the regression method, and it does not add other restrictions. Matching neither requires any functional form such as linearity on the outcome equations nor assumes homogeneous treatment effects across all the units. Both assumptions are barely confirmed by economic theory and data.

When one or two covariates are available, matching on covariate(s) is easy and accurate. When the number of covariates increases, matching directly on the covariates is impractical. Because the dimensions of covariates could be extremely large, good matches on all dimensions require a lot of computation, and sometime may not exist. This is called the curse of dimensionality. Rosenbaum and Rubin (1983) give an attractive way to overcome the curse of dimensionality, the propensity score matching method. However, the true propensity score is unknown in practical studies. To implement propensity score matching methods, we have to estimate propensity scores before matching. This paper focuses on different econometric models to estimate propensity scores. When the propensity score is unknown, it can be estimated by logit regression, probit regression, semiparametric single index regression and nonparametric regression. Logit regression and probit regression are easy to implement, and they will give good estimates when the data structure meets their assumptions on the functional form and the distribution of the error term. However, real data seldom meet those assumptions. To capture the true data structure, semiparametric regression and nonparametric regression can be employed. Nonparametric regression requires larger sample sizes than logit (probit) regression because the data must supply the model structure as well as the

model estimates. Semiparametric single-index regression involves an unknown finite-dimensional parameter and an unknown (link) function. It combines parametric and nonparametric models. The semiparametric single index model requires larger sample size than logit (probit) regression but smaller sample size than nonparametric regression. Both of semiparametric and nonparametric regressions have a slow convergence rate; they generally yield better fit of the data than parametric regressions when the data structure is not known.

However, the propensity score is just a tool to match observations. If the estimation of treatment effect is not affected by the models estimating propensity score, why bother do it nonparametrically rather than parametrically? Hahn (1998), Hirano, Imbens and Ridder (2003) show the true propensity score would not increase efficiency in estimating treatment effects. Moreover, many papers discuss the sensitivity of treatment effect to model specification of propensity score in a parametric context. Dehejia and Wahba (1999), Jalan and Ravallion (2003), Levine and Painter (2003), Heinrich, Meuser and Troske (2005) suggest that the specification of the propensity score is not important. Rubin and Thomas (1996) argue to include any meaningful variables in the propensity score model, while Bryson, Dorsett, and Purdon (2002) argue against including irrelevant variables on efficiency grounds. Brookhart et al. (2006) suggest that variables need to be related to the outcome to stay in propensity score model; including unnecessary variables would result in a higher mean squared error of the treatment effect. Smith and Todd (2005) evaluate the performance of propensity score matching estimators with the National Supported Work (NSW) Demonstration data and survey data from the Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID). Their

result shows that the propensity score specifications do impact results. Zhao (2008) investigates the sensitivity issue through Monte Carlo experiments. He finds that treatment effects estimated from the mis-specified models are almost as good as the ones from the correct models, as long as the matching assumptions are satisfied.

All of the above papers discussed sensitivity of treatment effect estimation to parametric specification, but they did not address the situation if we use semiparametric or nonparametric models to estimate propensity score. This paper uses Monte Carlo experiments to assess the properties of estimators of treatment effect under different propensity score estimation procedures: probit regression, logit regression, semiparametric singles index and nonparametric kernel methods.

The rest of the paper is organized as follows: Section 2 briefly reviews the treatment effect model using the potential outcome framework and outlines how to implement propensity score matching. Section 3 talks about four methods to estimated propensity score, probit, logit, semiparametric and nonparametric method respectively. Section 4 describes Monte Carlo experiments under the unconfoundedness assumption, and assesses estimators of the four regression methods. Section 5 concludes.

2.2. Treatment Effect Model and Propensity Score

The treatment effect model is used to estimate the effect of a binary treatment, T, where T=1 for an active treatment and T=0 for a control treatment. Let Y_1 denote a response variable that would be observed under the active treatment and Y_0 denote a response variable under the control treatment. The two variables are called potential

outcomes (Neyman 1923; Rubin, 1974, 1977). If we denote by X a vector of pre-treatment variables, each unit under study is sampled from the joint distribution of (Y, X, T), where$Y = TY_1 + (1 - T)\, Y_0$. Then we can define different treatment effects as below:

Average Treatment Effect on the Treated (ATT)

$$\gamma_{ATT} = E(Y_{i1} - Y_{i0} \,|T_i = 1)$$

Average Treatment Effect on the Control (ATC)

$$\gamma_{ATC} = E(Y_{i1} - Y_{i0} \,|T_i = 0)$$

Average Treatment Effect (ATE)

$$\gamma_{ATE} = E(Y_{i1} - Y_{i0}) = \gamma_{ATT} \Pr(T_i = 1) + \gamma_{ATC} \Pr(T_i = 0)$$

The average treatment effect on the treated (ATT) is also called causal effect; ATE and ATT attract the most attention in the literature. As the counterfactual outcomes $E(Y_{i1} \,|T_i = 0)$ and $E(Y_{i0} \,|T_i = 1)$ are not observed, proper substitutes for them have to be selected in order to estimate ATT and ATE. Using $E(Y_{i1} \,|T_i = 1)$ and $E(Y_{i0} \,|T_i = 0)$ for them is usually not a good idea, because it is most likely that factors, which determine the treatment decision, also affect the outcome variable of interest. Thus, Y_i and T_i are correlated, leading to a self-selection bias stated as below. For ATT it can be noted as:

$$E(Y_{i1} \,|T_i = 1) - E(Y_{i0} \,|T_i = 0) = \gamma_{ATT} + E(Y_{i0} \,|T_i = 1) - E(Y_{i0} \,|T_i = 0) \qquad (2.2.1)$$

The difference between the left hand side of equation (3.2) and γ_{ATT} is the so called "self-selection bias." The true parameter γ_{ATT} is only identified if

$$E(Y_{i0} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0) \quad = 0 \tag{2.2.2}$$

In randomized experiments, the treatment is assigned randomly, (2.2.2) is guaranteed, and the treatment effect is identified. In non-experimental studies, one has to impose some identifying assumptions to solve the problem stated in equation (2.2.1) as well as for estimation of ATE.

Assumption 1 (Unconfoundedness**)** Rosenbaum and Rubin (1983)

For each unit *i* with pre-treatment covariates X_i, $(Y_{i1}, Y_{i0}) \perp T_i \mid X_i$ (2.2.3)

X_i is the pre-treatment covariates for each unit, $\perp$ means independence. Thus, this assumption is also referred as "independence assumption." It implies that selection is solely based on observable characteristics X_i, and all variables that influence treatment assignment and potential outcomes simultaneously are observed by the researcher. This is a strong assumption but is fundamentally untestable.

Assumption 2 (Overlap)

$$0 < prob(T_i = 1 \mid X_i) < 1 \tag{2.2.4}$$

It ensures that units with the same X_i have a positive probability of being participants or non-participants. Rosenbaun and Rubin (1983) call the combination of the two assumptions as "strong ignorability."

With "strong ignorability", γ_{ATT} could be written as

$$\gamma_{ATT} = E_{X_i|T=1}\{E(Y_{i1}\,|T_i = 1, X_i) - E(Y_{i0}\,|T_i = 0, X_i)\,\} \tag{2.2.5}$$

One way to estimate this equation could be done by matching on their vector of covariates, X_i. In principle, we could stratify the data into sub-groups (or bins), each group defined by a particular value of X_i. Within each bin, we could calculate $Y_1 - Y_0$ conditioning on X_i. This method has a limitation, because it relies on a sufficiently rich comparison group so that no bin containing a treated unit is without a comparison unit. However, this is always impossible in reality. Rosenbaum and Rubin (1983) proposed propensity score matching to avoid it.

Assumption 1*(Unconfoundedness Based on Propensity Score)

Let $p(X_i)$ be the probability of a unit i having been assigned to treatment, defined as $p(X_i) = \Pr(T_i = 1 \mid X_i) = E(\,T_i|X_i)$. Then

$$(Y_{i1}, Y_{i0}) \perp T_i \mid X_i \implies (Y_{i1}, Y_{i0}) \perp T_i \mid p(X_i) \tag{2.2.6}$$

The conditional independence extends to the use of propensity score. Matching on the propensity score substantially reduces the dimensionality of matching on X_i, allowing us to match on a scalar rather than in a general n-space. With the assumption of $0 < prob(T_i = 1 \mid p(X_i)) < 1$, we can write (2.2.5) as

$$\gamma_{ATT} = E_{\mathrm{p}(X_i)\,|T=1}\{E\big(Y_{i1}\,\big|T_i = 1, \mathrm{p}(X_i)\big) - E\big(Y_{i0}\,\big|T_i = 0, \mathrm{p}(X_i)\big) \qquad (2.2.7)$$

Unbiased estimates of $E\big(Y_{i1}\,\big|T_i = 1, \mathrm{p}(X_i)\big)$ and $E\big(Y_{i0}\,\big|T_i = 0, \mathrm{p}(X_i)\big)$ could be obtained if $\mathrm{p}(X_i)$ is known.

It is important to note that covariates X_i and propensity score $\mathrm{p}(X_i)$ both belong to a class called balancing score. A balancing score is a function of the observed covariate X_i such that the conditional distribution of X_i is the same for the treated and control units (Rosenbaum and Rubin 1983). Covariate X_i is the finest balancing score, while the propensity score is the coarsest balancing score. There are infinitely many balancing scores between them. Theoretically, controlling for any balancing score is enough to correct the selection bias due to observables.

2.3. Estimations of Propensity Score

The propensity score is unknown in most cases; we need to estimate the propensity score beforehand to apply equation (2.2.7). To estimate the propensity score, we generally solve the following model and get $prob(T_i = 1 \mid X_i) = \mathrm{p}(X_i)$ as the estimated propensity score.

$$T_i = 1 \text{ If } T_i^* > 0$$

$$T_i = 0 \text{ Otherwise}$$

$$T_i^* = f(X_i'\beta) + \epsilon_i$$

where $f(\cdot)$ is unknown and ϵ_i is independent of X_i.

There are different models to solve the above problem. The easiest is using probit or logit model. However, probit or logit model require ϵ_i follow a normal or logit distribution, which always be satisfied in reality. In this case, semiparametric or nonparametric method could help to give a better fit. As propensity score is just the tool to match treated and control units, will the good fit of it affect treatment effect estimation? The rest of this paper will answer it. We illustrate the algorithms for parametric (logit and probit), semiparametric single index and nonparametric kernel methods to estimate propensity score, and did Monte Carlo experiment in the next section to check how sensitive the treatment effect estimate to these four methods.

2.3.1. Parametric Method to Estimate Propensity Score

For ease of estimation, most applications in the statistics literature have concentrated on logit (or probit) model. Both logit and probit model assume that $f(\cdot)$ has a linear functional form, but they use different assumptions for the error term. Logit model assumes ϵ has a symmetric logistic distribution, in this case,

$$\mathrm{p}(X_i) = E(T_i \,|X_i) = \frac{1}{1+e^{-X_i'\beta}} \qquad (2.3.1)$$

Probit model assumes $\epsilon \sim N(0, \sigma^2)$, in this case,

$$\mathrm{p}(X_i) = E(T_i \,|X_i) = \phi(X_i'\beta) \tag{2.3.2}$$

Here, T_i is the treatment status indicating if the unit receives treatment, and $\phi(\cdot)$ is the cumulative distribution function (CDF) of standard normal distribution.

The above statement tells that different distributional assumptions for ϵ_i lead to different functional forms for the conditional probability of $T = 1$. Hence, the consistency of estimation requires the correct distributional specification of ϵ_i. Additionally, we have no reason to believe that the logit (or probit) model could capture the underlying nonlinear pattern in the propensity score estimation, which is another possibility to yield inconsistent estimates or poor predictions.

2.3.2. Semiparametric Single-Index Method to Estimate Propensity Score

These semiparametric single index methods requires the $E(T|X)$ depends on the vector X through a single linear combination $X'\ \beta$.

The semiparametric single index regression model is

$$\mathrm{p}(X_i) = \ E(T_i \,|X_i'\beta) = g(X_i'\beta).$$

$\beta \in \mathbb{R}^k$ (k>1 is the dimension of X_i) and $g\colon \mathbb{R} \to \mathbb{R}$ are unknown. $g(\cdot)$ is sometimes called a link function. If X_i is one-dimensional, k=1. The model is the one-dimensional nonparametric regression with no semiparametric component.

Identification of β and g require that X_i contains at least one continuously distributed variable, and that this variable has a non-zero coefficient, because it is impossible to identify a continuous function fully on discrete support.

In the case of an unknown function $g(\cdot)$, the kernel method cannot estimate $g(X_i'\beta)$ directly, because both $g(\cdot)$ and β is unknown. But for a given value $\tilde{\beta}$, we could estimate

$$G(X_i'\tilde{\beta}) \stackrel{\text{def}}{=} E(T_i|X_i'\tilde{\beta}) = E(g(X_i'\beta)|X_i'\tilde{\beta})$$

by kernel method, and the last equality follows from the fact that ϵ_i and X_i are independent.

When $\beta = \tilde{\beta}$, $G(X_i'\tilde{\beta}) = g(X_i'\tilde{\beta})$, but in general $G(X_i'\tilde{\beta}) \neq g(X_i'\tilde{\beta})$, if $\beta \neq \tilde{\beta}$. A leave-one-out nonparametric kernel estimator of $G(X_i'\tilde{\beta})$ is given by

$$\hat{G}_{-1}(X_i'\tilde{\beta}) \equiv \hat{E}_{-1}(T_i|X_i'\tilde{\beta}) = \frac{\sum_{j=1,j\neq i}^{n} T_i K(\frac{X_j'\tilde{\beta} - X_i'\tilde{\beta}}{h})}{\sum_{j=1,j\neq i}^{n} K(\frac{X_j'\tilde{\beta} - X_i'\tilde{\beta}}{h})} \quad (2.3.3)$$

Klein and Spady (1993) suggested estimating β by maximum likelihood methods. The estimated log-likelihood function is

$$l(\beta) = \sum_i \left\{(1-T_i)\ln\left(1-\hat{g}_{-1}(X_i'\tilde{\beta})\right) + T_i \ln\left(\hat{g}_{-1}(X_i'\tilde{\beta})\right)\right\} \quad (2.3.4)$$

where $\hat{g}_{-1}(X_i'\tilde{\beta})$ is defined as (2.3.3). Maximizing the likelihood function with respect to β leads to the semiparametric maximum likelihood estimator of β. Klein and Spady

showed this estimate is $\sqrt{n}$ consistent and semiparametrically efficient. And it has asymptotic normal distribution.

2.3.3. Nonparametric Kernel Estimation of Propensity Score

Nonparametric estimation of propensity score has no assumption on ϵ_i or the relationship between ϵ_i and X_i . Li, Racine and Wooldrige (2009) proposed nonparametric kernel estimation for this case.

They use X_i^c and X_i^d to denote the continuous and discrete components in X_i. Let $\omega(\cdot)$ denote a univariate kernel function for the continuous variable, and define the product kernel function of the continuous variables by

$$W\left(X_i^c,\ X_j^c, h\right) = \prod_{s=1}^{q} h_s^{-1}\, \omega\left(\frac{X_{is}^c - X_{js}^c}{h_s}\right) \tag{2.3.5}$$

where X_{is}^c is the s-th component of X_i^c and h_s is the corresponding smoothing parameter (s=1,…q). For the discrete variables, divide them into two sets, one set contains discrete variables with natural ordering, the other set contains discrete variables that do not have a natural ordering. The kernel for ordered discrete variables is $l_o\left(X_{is}^d,\ X_{js}^d,\ \lambda_s\right)$, while $l_u\left(X_{is}^d,\ X_{js}^d,\ \lambda_s\right)$ is for the unordered variables.

$$l_o\left(X_{is}^d, X_{js}^d, \lambda_s\right) = \begin{cases} 1 & X_{is}^d = X_{js}^d \\ \lambda_s^{|X_{is}^d - X_{js}^d|} & X_{is}^d \neq X_{js}^d \end{cases}$$

$$l_u\left(X_{is}^d, X_{js}^d, \lambda_s\right) = \begin{cases} 1 & X_{is}^d = X_{js}^d \\ \lambda_s & X_{is}^d \neq X_{js}^d \end{cases}$$

Combining $l_o(X_{is}^d, X_{js}^d, \lambda_s)$ and $l_u(X_{is}^d, X_{js}^d, \lambda_s)$ together, we could derive the product kernel function for categorical variables as below:

$$L(X_i^d, X_j^d, \lambda)=\left[\prod_{s\in S_o} \lambda_s^{|X_{is}^d - X_{js}^d|}\right]\left[\prod_{s\in S_u} \lambda_s^{\mathbf{1}(X_{is}^d \neq X_{js}^d)}\right] \quad (2.3.6)$$

where S_o and S_u denote the index sets for ordered and unordered components of X_i^d; $\mathbf{1}(\cdot)$ is an indicator function. Taking product of (2.3.5) and (2.3.6), we could obtain the kernel function for a mixture of categorical and continuous variables.

$$K_{ij,\alpha} = K(X_i, X_i, \alpha) = W(X_i^c, X_j^c, h) \cdot L(X_i^d, X_j^d, \lambda) \quad (2.3.7)$$

where $\alpha = (h, \lambda)$, $W(\cdot)$ and $L(\cdot)$ are the kernel functions defined in (2.3.5) and (2.3.6)

Using $K_{ij,\alpha}$, the estimation of propensity score $P(X_i) = Prob(T_i = 1 \mid X_i) = E(T_i \mid X_i)$ is

$$P(X_i) = \frac{\sum_{j=1}^n T_j K_{ij,\alpha}}{\sum_{j=1}^n K_{ij,\alpha}} \quad (2.3.8)$$

Smoothing parameters $\alpha = (h, \lambda)$ could be chosen by minimizing leave-one out least square cross validation.

No matter which method is used, we will have estimated propensity scores, and ATT is calculated by matching in the second stage.

2.4. Monte Carlo Study

Following Zhao (2008), Monte Carlo experiment is used to examine the sensitivity to different propensity score estimation methods based on the potential outcome model. Artificial samples are generated according to the following rule".

$Y_{i1} = X_i'\alpha_1 + v_{i1}$ Outcome in treated state

$Y_{i0} = X_i'\alpha_0 + v_{i0}$ Outcome in untreated state

$Y_i = T_i Y_{i1} + (1 - T_i)Y_{i0}$ Observed outcome

$T_i^* = f(X_i) + \epsilon_i$ Latent index function

$T_i = I(T_i^* > 0)$ Treatment indicator

where $I(\cdot)$ is an indicator function. v_{i1} and v_{i0} are drawn from standard normal distribution. T_i is a binary variable, and Y_{i1}, Y_{i0} take linear functional form of X_i. The distribution of ϵ_i, together with the value of X_i, determines the number of treated and untreated observations in the sample.

The idea of this experiment is to simulate T_i^* and T_i using different assumptions of the distribution of the error term ϵ_i. We consider four distributions of ϵ_i (normal, logistic, bimodal and heteroscedastic distributions). We use probit, logit, semiparametric single index model, and nonparametric regression to estimate propensity score. Thus, the propensity scores are estimated from mis-specified model in some cases. With the estimated propensity score, we carry on treatment effect calculation based on caliper

matching. As we generate the data, the true treatment effects could be calculated, we can assess the sensitivity of treatment effect to different estimation models for propensity score by examining the closeness of estimated treatment effects to the true values.

Outcomes for Monte Carlo experiment are based on 200 replications, and sample size of 1000 is used for each replication. In the following two Monte Carlo experiments, we assume T_i^* takes linear form of X_i in experiment I, and nonlinear form for experiment II.

2.4.1. Monte Carlo Experiment I

Suppose T_i^* takes a linear form of X_i,

$$T_i^* = X_i'\beta + \epsilon_i$$

Four distributions of error term are considered.

(1) $\epsilon_i \sim N(0,1)$

(2) $\epsilon_i \sim Logistic$

(3) $\epsilon_i \sim 50 - 50\ mixture\ of\ two\ normal\ dist.\ N(3,1) and\ N(-3,1)$

(4) $\epsilon_i = 0.25[1 + 2(X_i'\beta)^2 + (X_i'\beta)^4]u_i\,, u_i \sim Logistic$

Parametric probit mode is the correct model for assumption (1), and logit model is the correct model for assumption (2). Assumption (3) supposes the error term takes a bimodal distribution and assumption (4) gives a heteroscedastic error term. Semiparametric single index model and nonparametric model will give better fits than probit and logit model for the two cases.

Correlation between error terms v_{i1} and v_{i0} in the outcome equations is allowed, but $cov(v_{i1},\ \epsilon_i) = 0$ and $(v_{i0},\ \epsilon_i) = 0$. Thus, the unconfoundedness assumption is satisfied. The four distributions of ϵ_i make the number of treated units account for 50.0%, 69.3% 50.0% and 70.9% of the sample respectively.

Table 2.1 Monte Carlo Experiment I: Linear Index function; Matching with replacement

	probit			logit		
	bias	standard error	mse	bias	standard error	mse
True Value	0.000	N/A	N/A	0.000	N/A	N/A
Normal	0.316	0.111	0.113	0.301	0.109	0.103
Logistic	0.395	0.306	0.251	0.395	0.292	0.243
Bimodal	0.133	0.071	0.023	0.133	0.069	0.023
Heteroscedastic	-0.189	0.063	0.063	-0.145	0.142	0.042
	semiparametric			nonparametric		
	bias	standard error	mse	bias	standard error	mse
True Value	0.000	N/A	N/A	0.000	N/A	N/A
Normal	0.287	0.143	0.104	0.330	0.133	0.128
Logistic	0.388	0.307	0.246	0.493	0.359	0.374
Bimodal	0.129	0.085	0.024	0.125	0.090	0.024
Heteroscedastic	0.249	0.306	0.156	0.305	0.353	0.219

Table 2.1 presents the statistics of estimated ATT by radius matching with replacement. We report the bias, the standard deviation and MSE for estimated ATT normalized by true ATT (estimated ATT/true ATT). The bias for true value is of course

equal to zero. Comparing the bias calculated for the four different regression methods, we could tell that the ATTs gotten from mis-specified models are comparable to the ones from the correct models, semiparametic single index model reduces the bias a bit for the cases of normal, logistic and bimodal error terms, but the bias is still close to the biases of probit and logit models. Nonparametric regression method is not favored in the cases of normal, logistic and heteroscedastic error terms, it produces the smallest bias for the bimodal error term, but the improvement is very small. The bias of logit model is consistently smaller than that for probit model, no matter which error term specification is used, but the magnitude is close to zero, so logit model is marginally better than probit model. Although biases of the four regression methods are different, they are very close, so we claim that estimation methods of propensity score have very limited influence on the estimation of ATT.

Table 2.2 gives the estimates by matching without replacement; we could see an increase in bias for majority cases, compared to matching with replacement. So matching with replacement is preferred. Again the four different estimation methods give comparable estimates, and logit model is marginally better than probit model.

Table 2.2 Monte Carlo Experiment I: Linear index function; Matching without replacement

	probit			logit		
	bias	standard error	mse	bias	standard error	mse
True Value	0.000	N/A	N/A	0.000	N/A	N/A
Normal	0.373	0.142	0.161	0.364	0.142	0.154
Logistic	0.338	0.363	0.247	0.337	0.300	0.205
Bimodal	0.146	0.078	0.028	0.144	0.088	0.029
Heteroscedastic	0.035	0.143	0.022	0.071	0.131	0.022
	semiparametric			nonparametric		
	bias	standard error	mse	bias	standard error	mse
True Value	0.000	N/A	N/A	0.000	N/A	N/A
Normal	0.367	0.173	0.166	0.334	0.171	0.156
Logistic	0.356	0.363	0.260	0.504	0.444	0.499
Bimodal	0.131	0.095	0.026	0.119	0.119	0.031
Heteroscedastic	0.410	0.272	0.244	0.333	0.440	0.336

2.4.2. Monte Carlo Experiment II

Now let's consider the case that T_i^* takes a nonlinear form of X_i, $T_i^* = f(X_i) + \epsilon_i$. The four distributions of the error terms, considered in Monte Carlo Experiment I, are considered here as well and we assume $cov(v_{i1}, \epsilon_i) = 0$ and $cov(v_{i0}, \epsilon_i) = 0$ to satisfy the unconfoundedness assumption. The four distributions of ϵ_i make the number of treated units account for 37.3%, 43.3% 47.0% and 48.6 % of the sample respectively.

Table 2.3 reports the Monte Carlo experiment results when matching are carried out with replacement and Table 2.4 are the results without replacement. As $f(X_i)$ takes nonlinear form, logit model, probit model and semiparametric single index regression model are all mis-specified models for the data, nonparametric regression will estimate the propensity score closer to the true propensity score. The results in table 2.3 show that bias of estimated ATT with probit or logit model is much smaller than that of the nonparametric regression model, while semiparametric single index regression method gives comparable bias in some cases, but performs worse when the error term has heteroscedasticity problem.

Table 2.3 Monte Carlo Experiment II: Nonlinear index function; Matching with replacement

	Probit			logit		
	bias	standard error	mse	bias	standard error	mse
True Value	0.000	N/A	N/A	0.000	N/A	N/A
Normal	-0.074	0.188	0.040	-0.050	0.191	0.039
Logistic	-0.163	0.157	0.051	-0.122	0.131	0.032
Bimodal	0.167	0.271	0.100	0.165	0.275	0.102
Heteroscedastic	-0.093	0.068	0.013	-0.060	0.064	0.008
	semiparametric			nonparametric		
	bias	standard error	mse	bias	standard error	mse
True Value	0.000	N/A	N/A	0.000	N/A	N/A
Normal	0.073	0.224	0.055	1.279	0.109	1.648
Logistic	0.009	0.173	0.029	1.018	0.101	1.046
Bimodal	0.282	0.228	0.131	0.930	0.083	0.872
Heteroscedastic	0.366	0.127	0.150	1.160	0.059	1.349

From Table 2.4, we could notice that the bias increase substantially without replacement for nonlinear index function. Not only did the nonparametric regression method do a poor job, but also probit, logit and semiparametric single index regression methods. As the case of linear index function, matching without replacement is not recommended.

Table 2.4 Monte Carlo Experiment II: Nonlinear index function; Matching without replacement

	Probit			logit		
	bias	standard error	mse	bias	standard error	mse
True Value	0.000	N/A	N/A	0.000	N/A	N/A
Normal	1.146	0.095	1.322	1.150	0.092	1.332
Logistic	1.032	0.068	1.071	1.032	0.068	1.070
Bimodal	0.837	0.069	0.706	0.839	0.069	0.709
Heteroscedastic	1.109	0.063	1.233	1.109	0.063	1.233
	semiparametric			nonparametric		
	bias	standard error	mse	bias	standard error	mse
True Value	0.000	N/A	N/A	0.000	N/A	N/A
Normal	1.122	0.093	1.268	1.281	0.158	2.083
Logistic	0.942	0.086	0.894	1.021	0.152	1.332
Bimodal	0.832	0.076	0.698	0.929	0.147	1.106
Heteroscedastic	1.112	0.062	1.240	1.163	0.091	1.700

In summary, if the index function is in linear form, results of Monte Carlo experiments suggest that ATT estimator is insensitive to the estimation methods of propensity scores. When index function is nonlinear, probit and logit model do better jobs than semiparametric single index model and nonparametric regression model. Logit model does a slightly a better job than probit model. As long as semiparametric single index regression and nonparametric regression cannot bring improvements in estimates of ATT and they normally require high computation load, we suggest using probit and logit model to compute propensity score. However, the simulation exercises in this chapter only cover a limited number of cases; we need to be cautious when generalizing the findings here to other scenarios.

2.5. Conclusion

Probit mode, logit model, semiparametric single index model and nonparametric regression model are possible ways to estimate propensity score. We use Monte Carlo simulation to examine the sensitivity of ATT estimates to these four methods under the unconfoundedness condition. The simulation results show that ATT estimates are insensitive to estimation methods when the index function for treatment is in linear form. When the index function is in nonlinear form, probit model and logit model give less bias than the nonparametric regression model, so nonparametric regression model is not recommended. Semiparametric single index model sometimes give slightly less bias than probit and logit model, but this trend is not consistent. Because the magnitude of

reduction in bias is very small and semiparametric model require a lot more computational load, we recommend probit and logit models.

Chapter 3

Use of Cross-Sectionally Varying Coefficient Models to Account for Heterogeneous Treatment Effects in Nonexperimental Studies

3.1. Introduction

Many articles have been written estimating the average treatment effect of a treatment on the treated (ATT) or the average treatment effect (ATE). Rosenbaum (1995), Robins and Rotnizky (1995), Hahn (1998), Heckman, Ichimura and Todd (1997, 1998), Imbens, Newey and Ridder (2003), Hirano, Imbens and Ridder (2003) and others have used various methods to estimate ATT and ATE in the nonrandomized treatment assignment setting. These approaches tend to focus exclusively on the mean impact, the

first moment of treatment effect. They are elaborately reviewed by Imbens (2004), Heckman and Vytlacil (2005), and Imbens and Wooldridge (2009).

The methods that focus on ATT and ATE, either employing matching, regression or instrumental method, have problems. Simple matching estimators include a conditional bias term of stochastic order $N^{-1/k}$, where k is the number of continuous matching variables, and thus the matching estimators are in general not $N^{1/2}$ consistent when more than two continuous covariates are used for matching. Moreover, the estimators with a fixed number of matches do not achieve the semiparametric efficiency bound. Regression methods suffer from specification error. For instrumental variables method, it can be challenging to find an instrument that is both relevant (not weak) and exogenous, and assessment of instrument exogeneity can be highly subjective. Other than these problems, the ATT and ATE provide rather limited information on the value of the first moment. If and only if the treatment has a homogeneous impact on each observation could the first moment provide a treatment effect for an individual, otherwise it only provides the average level of a treatment's overall impact. However, a homogeneous impact implies that the treatment has the same effect on each unit, which is unlikely, as people or treated units are not identical and may respond differently to the same treatment. Since assuming equal impact on each unit is implausible, we should relax it to allow for a heterogeneous treatment effect. Some researchers have developed new methods to account for heterogeneous treatment effects and estimate their distribution. Extending classical probability theory in the framework of treatment effect, Heckman, Smith, and Clements (1997) first provide robust evidence that heterogeneous treatment effects exist. Abadie,

Angrist, and Imbens (2002) use instrumental variable methods to identify the quantile treatment effect. Chernozhukov and Hansen (2005) impose restrictions on the evolution of ranks across treatment states and employed quantile regression methods to recover the causal effects of treatment on the quantiles for the outcomes of economic variables of interest. Firpo (2007) use semiparametric methods to compute the quantiles of the marginal distribution of potential outcomes, an example of which is the value a response variable would take for an individual had the treatment, denoted by T, been the alternative value t. This is referred to as the counterfactual event. Firpo propose that simple differences in potential outcome quantiles could be used as the treatment effects at quantiles, if rank preservation holds. Athey and Imbens (2006) focus on the effects at quantiles in the situation in which repeated cross sections or longitudinal data are available, and present what they call "Change-in-Change method" to estimate the entire counterfactual distribution nonparametrically. These papers give treatment effect on quantiles of outcome rather than the distribution of treatment effect. Wu and Perloff (2006) approximate the distribution of individual treatment effects using the deconvolution method. The deconvolution method could only estimate the distribution when units are randomly assigned into treatment, but it does not take account of the case in which units are self-selected into treatment. Fan and Park (2009) study partial identification of the distribution of treatment effects of a binary treatment for ideal randomized experiments, ideal randomized experiments with a known value of a dependence measure and for data satisfying the selection-on-observables assumption respectively. Fan and Wu (2010) establish sharp bounds on the joint distribution of potential outcomes and the distribution of treatment effects in switching regime models.

The methods mentioned above, either approximating treatment effects on quantiles of outcome or the distribution of treatment effects, do not in general identify individual treatment effect.

We present and develop a novel method of estimating individual treatment effects under the condition that units are self-selected into treatment. As long as the individual treatment effect is estimated, the impact distribution (or treatment effect distribution), the average treatment effect on the treated population, and the treatment effect on the whole population can be calculated. Because the observed outcome variable y is either the outcome under treatment y_1 or outcome that is not treated y_0, y could be written as $y = (y_1 - y_0)T + y_0$, where T indexes whether treated or not, $y_1 - y_0$ is the individual treatment effect, unknown for each unit, and y_0 is the unknown counterfactual outcome for treated units. In this paper, We reparameterize the terms $y_1 - y_0$, and y_0, express them in terms of a set of observable variables. Thus estimation of $y_1 - y_0$ and y_0 for each individual is transformed into the estimation of parameters for the observable variables, whose number is much smaller. We use a consistent and efficient Bayes estimator for the parameters of the observable variables, with which $y_1 - y_0$ and y_0 can be derived. The novelty of the paper's method is that it does not require either completely random treatment assignment or data on pairs of individuals matched by some criterion, one subjected to no treatment and the other subjected to the treatment. Moreover, it provides a way to reveal individual treatment effects.

The remainder of the paper is divided into five sections. Section 2 considers the observed outcome of a treatment in terms of the treatment dummy variable taking the

value 1 for the treated and 0 for the untreated individuals. The model is further parameterized making its coefficients functions of certain observable variables (defined as "coefficient drivers" in Section 2.2). Assuming that the regressors of the model are conditionally independent of its coefficients given the coefficient drivers, the conditional mean, variance and distribution of observable outcomes given the coefficient drivers can be derived. Identification and a method of estimation of the coefficient drivers' parameters for the model are discussed in the section as well. The coefficient drivers and estimated parameters together give the estimates of treatment effects. We call this the Cross-Sectionally Varying Coefficient model. The design and results of a simulation experiment are presented in Section 3. In section 4, the proposed method is applied in two studies. One examining the impact of the job training program on earnings based on the National Job Training Partnership Act (JTPA) Study, while the other examining the impact of seasoned equity offerings on the operational performances of firms, the operating income before depreciation and amortization over asset (OIBD/asset) specifically, using the merged data of Securities Data Corporation (SDC) New Issues database, Center for Research in Securities Prices (CRSP), and COMPUSTAT. Section 5 concludes.

3.2. General Modeling Framework for Estimating the Treatment Effects in Non-experimental Situations

3.2.1. Model Structure

We begin with N units, indexed by $i = 1,\ldots, N$, which are assumed to be drawn randomly from a large population. Let T_i denote the treatment dummy variable taking the value 1 if the i-th unit is treated and 0 if the i-th unit is not treated. The value $T_i = 1$ represents an active treatment and the value $T_i = 0$ a control treatment. Individual i shows a response, denoted by y_{i1}, to treatment if $T_i = 1$ and shows a response, denoted by y_{i0}, to no treatment if $T_i = 0$, $i = 1, \ldots, N$. One complication we have to deal with is that y_{i0} and y_{i1} cannot be observed for an individual at the same time. For untreated individual i, y_{i0} is an outcome in the absence of treatment and y_{i1} is an unrealized observation on what the effect on the individual would have been had the individual been treated. For treated individual i, y_{i1} is an observed outcome and y_{i0} is an unrealized observation on what the outcome would have been had the individual not been treated. In Neyman's (1923) and Rubin's (1974, 1977) terminology, y_{i0} is a potential outcome if individual i is treated and y_{i1} is a potential outcome if individual i is untreated. For each individual i, the observed outcome is

$$\begin{aligned} y_i &= T_i\, y_{i1} + (1 - T_i)\, y_{i0} \\ &= y_{i0} + (y_{i1} - y_{i0})\, T_i = \beta_{i0} + \beta_{i1}\, T_i = (1,\, T_i)\, \boldsymbol{\beta}_i \end{aligned} \qquad (3.2.1)$$

where $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1})'$, $\beta_{i0} = y_{i0}$, and $\beta_{i1} = (y_{i1} - y_{i0})$. These definitions imply that the determinants of β_{i0} are the same as those of y_{i0} and the determinants of β_{i1} are the same as those of $(y_{i1} - y_{i0})$. Because the coefficient $\boldsymbol{\beta}_i$ differs among units, we call the model Cross-Sectionally Varying Coefficient model (CVC model in short), and $\boldsymbol{\beta}_i$ CVC estimator. We have observations on y_i and T_i for *N* individuals to be denoted by *i* = 1, …, *N*.

Let X denote a vector of the determinants of y_{i0} and y_{i1}. It is important that these variables are not affected by the treatment. Often they take their values prior to the unit being exposed to the treatment. Each individual unit under study is sampled from the joint distribution of (*Y, X, T*), which refers to the distribution induced by the random sampling from the super population; $Y = \mathrm{T}Y_1 + (1 - T)\, Y_0$; $Y_1 = h_1(X) + \upsilon_1$; and $Y_0 = h_0(X) + \upsilon_0$, where h_0 and h_1 may not be linear functions. In our simulation experiment below, we work with three different pairs of functions (h_0, h_1) and the standard normal distribution for the pair (υ_0, υ_1) of errors are used.

Econometric textbooks discuss the method of instrumental variables (IV) for the treatment effects model. This method cannot be applied to model (3.2.1) because Swamy and Hall (2011) proved that such variables do not exist. The validity of this proof depends on the uniqueness of the coefficients and error terms of the models Swamy and Hall use in their proof. Any proof of the existence of IV based on the models with nonunique coefficients and error term is invalid. To solve the model, the following key assumption is made about treatment assignment.

Unconfoundedness assumption:

$$(y_{i0}, y_{i1}) \perp T_i \mid \boldsymbol{x_i} \tag{3.2.2}$$

where $\perp$ means independence. Equation (3.2.2) implies that (y_{i0}, y_{i1}) , with one of them being a potential outcome, are conditionally independent of T_i given $\boldsymbol{x_i}$. Thus by adjusting for differences in observed pretreatment variables $\boldsymbol{x_i}$ one can remove biases from comparisons between treated and control units. For example, a job training program is a typical case for treatment effect analysis and researchers have claimed that the training program is not randomly assigned to everyone because people have different tendencies to enroll in a program, which also impact the outcome thus inducing a so-called 'self-selection bias'. The unconfoundedness assumption provides a way around this problem by assuming that the program is randomly assigned among people who have the same characteristics (age, gender, education, marital status, experience and etc.), and thus a comparison between the treated unit and control unit with the same characteristics yields the unbiased treatment effect. This assumption was proposed by Rosenbaum and Rubin (1983), who refer to it as "ignorable treatment assignment." In the literature, different versions of this assumption are referred to as the endogeneity assumption, selection on observables or conditional independence. In our case specifically, this unconfoundedness assumption is equivalent to independence of T_i and v_i conditional on $\boldsymbol{x_i}$, as y_i is an additive function of $h(x_i)$ and v_i.

Assumption (3.2.2) requires that conditional on observed covariates there are no unobserved factors that are associated both with the assignment and with the potential

outcomes. We consider an assumption that is similar to assumption (3.2.2) in the next section.

3.2.2. Parameterizing Model

The assumptions we make are

Assumption I: The coefficient vector $\boldsymbol{\beta}_i$ in equation (3.2.1) can be linearly represented as

$$\boldsymbol{\beta}_i = \Pi \boldsymbol{d}_i + \boldsymbol{\varsigma}_i \tag{3.2.3}$$

where $\boldsymbol{\varsigma}_i = (\varsigma_{i0}, \varsigma_{i1})'$ is i.i.d. multi-normal, and $E(\boldsymbol{\varsigma}_i \mid \boldsymbol{d}_i) = 0$, $E\left(\boldsymbol{\varsigma}_i \boldsymbol{\varsigma}_i' \middle| \boldsymbol{d}_i\right) = E(\boldsymbol{\varsigma}_i \boldsymbol{\varsigma}_i') = \Delta$ for $\forall i$; $\boldsymbol{d}_i$ is a p-vector of observable variables explaining the cross-sectional variation in $\boldsymbol{\beta}_i$, these variables moving the coefficients are defined as coefficient drivers, and Π is a $2 \times p$ matrix of fixed parameters, which could transform the coefficient drivers to $\boldsymbol{\beta}_i$ if the values were known. Our problem of identifying $\beta_{i0} = y_{i0}$, and $\beta_{i1} = (y_{i1} - y_{i0})$ for all of the N units are simplified to the identification of Π, which has only $2p$ unknowns.

Substituting equation (3.2.3) into equation (3.2.1) gives

$$y_i = (\boldsymbol{d}_i' \otimes (1, T_i))\boldsymbol{\pi}^{Long} + (1, T_i)\boldsymbol{\varsigma}_i \tag{3.2.4}$$

where $\otimes$ denotes a Kronecker product and $\boldsymbol{\pi}^{Long}$ is a $2p$-vector given by a column stack of Π. It follows from model (3.2.4) that one of the uses of coefficient drivers is to parameterize model (3.2.1). Model (3.2.4) has the capability of estimating the cross-

sectionally varying coefficients $\boldsymbol{\beta}_i$. We call this model "a cross-sectionally varying coefficient model with the vector $\boldsymbol{d}_i$ of coefficient drivers (CVC$\boldsymbol{d}_i$)."

Another fundamental assumption we need to impose is

Assumption II: For $i = 1, \ldots, N$: $\boldsymbol{\beta}_i \perp T_i \mid \boldsymbol{d}_i$

Thus, a second use of coefficient drivers is to make $\boldsymbol{\beta}_i$ conditionally independent of T_i given $\boldsymbol{d}_i$. This is an extended form based on unconfoundedness assumption, equation (3.2.2). From equation (3.2.2), it is easy to derive that y_{i0} , $(y_{i1} - y_{i0}) \perp T_i \mid \boldsymbol{x}_i$, which is exactly $\boldsymbol{\beta}_i \perp T_i \mid \boldsymbol{x}_i$. Assumption II here is different from equation (2.2) in that it conditions on coefficient drivers $\boldsymbol{d}_i$ rather than the observable variables $\boldsymbol{x}_i$. In fact, $\boldsymbol{x}_i$ are good candidates for $\boldsymbol{d}_i$. We are fine to make $\boldsymbol{d}_i = \boldsymbol{x}_i$, if $\boldsymbol{\beta}_i$ takes linear form of $\boldsymbol{x}_i$. However, $\boldsymbol{\beta}_i$ more often takes nonlinear form of $\boldsymbol{x}_i$, so $\boldsymbol{d}_i$ cannot simply be the same as $\boldsymbol{x}_i$. $\boldsymbol{d}_i$ has to include $\boldsymbol{x}_i$ plus some functional forms of $\boldsymbol{x}_i$, because Assumption I require $\boldsymbol{\beta}_i$ takes linear form of $\boldsymbol{d}_i$, $\boldsymbol{d}_i$ should absorb the nonlinear part into itself. In empirical analysis, the functional form of $\boldsymbol{\beta}_i$ in terms of $\boldsymbol{x}_i$ is never known, we could include squared terms, cross-product terms, log terms and many functional form we could think up in our $\boldsymbol{d}_i$, then it comes to the pervasive variable selection problem in statistical application. We suggest following theoretical models to pick $\boldsymbol{x}_i$ and functions of $\boldsymbol{x}_i$, fitness of estimated y and true y is also a good criterion to refer.

Intuitively, Assumption II states that selection into the treatment is random (i.e., exogenous) conditional on a set of coefficient drivers. This assumption also serves as criteria for selection of coefficient drivers.

Assembling the data we have on each variable in (2.4) in a matrix gives

$$\boldsymbol{y} = A\,\boldsymbol{\pi}^{Long} + \mathrm{D}\boldsymbol{\varsigma} \tag{3.2.5}$$

where $\boldsymbol{y} = (y_1,...,y_N)'$ is an N-vector; $A = (\boldsymbol{d}_1 \otimes (1,T_1)',\ \ldots,\ \boldsymbol{d}_N \otimes (1,T_N)')'$ is $N \times 2p$; $\mathrm{D} = diag_{1\le i\le N}((1,T_i))$ is $N \times 2N$; and $\boldsymbol{\varsigma} = (\varsigma_1',...,\varsigma_N')'$ is a $2N$-vector, and $E(\boldsymbol{\varsigma}\,\boldsymbol{\varsigma}') = I_N \otimes \Delta$ according to assumption I.

Assumption III: The matrix A in (3.2.5) has full column rank and D has full row rank.

It is a fact that D has full row rank, and A would has full column rank as long as each unit is independent. With Assumption III, the vectors $\boldsymbol{\pi}^{Long}$ and $\mathrm{D}\boldsymbol{\varsigma}$ are identifiable.

Under Assumption I-III with Assumption IV (IV') listed below, Swamy Yaghi, Mehta and Chang (2007) derive the formulae for best linear unbiased predictor (BLUP) for $\boldsymbol{\pi}^{Long}$ and $\mathrm{D}\boldsymbol{\varsigma}$, they further provided the estimators' consistent property. Swamy and Mehta (1975), Swamy and Mehta (1976), and Swamy and Tinsley (1980) give detailed derivation and proof about consistency and asymptotic distribution in general scenario. For our case specific for treatment effect, where $y_i = (1,\ T_i)\,\boldsymbol{\beta}_i$, we offered a succinct and transparent way to derive $\boldsymbol{\pi}^{Long}$ and prove its asymptotic properties as below. If you

are interested in proof when explanatory variables for y are more than two dimensions, ς_i is not i.i.d. and etc., please refer Swamy and Mehta (1975) and Swamy and Mehta (1976).

3.2.3. A Simple Bayes Estimator for the Cross-sectionally Varying Coefficient Model with Coefficient Drivers

Assumption IV: The coefficient vector $\boldsymbol{\pi}^{Long}$ is a priori distributed as normal with mean vector $\boldsymbol{\gamma}$ and covariance matrix Ω (Ω is not singular), i.e.

$$\pi^{Long} \sim N(\gamma, \Omega) \tag{3.2.6}$$

For notational simplicity, we denote $Z = \{d_1, \dots, d_N, T_1, \dots, T_N\}$, if Z is given, $A = (\boldsymbol{d}_1 \otimes (1, T_1)', \dots, \boldsymbol{d}_N \otimes (1, T_N)')'$ is known. Assumption I and II allow us to construct the following moments:

$$E(y|Z, \pi^{Long}) = E\left(A\pi^{Long} + \mathrm{D}\varsigma \,\middle|\, Z, \pi^{Long}\right) = A\pi^{Long} + DE\left(\varsigma \,\middle|\, Z, \pi^{Long}\right) = A\pi^{Long}$$

$$cov(y|Z, \pi^{Long}) = E\left\{\left(y - E(y|Z, \pi^{Long})\right)\left(y - E(y|Z, \pi^{Long})\right)' \middle| Z, \pi^{Long}\right\}$$

$= E\{\mathrm{D}\varsigma(\mathrm{D}\varsigma)'|Z, \pi^{Long}\} = \mathrm{DE}(\varsigma \quad \varsigma'|\, Z, \pi^{Long})D'$

$= \mathrm{DE}(\varsigma \quad \varsigma')D' = \mathrm{D}(I_N \otimes \Delta)D'$.

To conclude, under Assumptions I-IV, model (2.5) implies that

$$y|Z, \pi^{Long} \sim N(A\pi^{Long}, W), \ where \ W = \mathrm{D}(I_N \otimes \Delta)D' \tag{3.2.7}$$

The posterior for π^{Long} thus becomes

$$p(\pi^{Long}|y, Z) \quad \propto \quad p(y|\pi^{Long}, Z) \times p(\pi^{Long})$$

$$\propto \exp\left\{-\frac{1}{2}[(y - A\pi^{Long})'W^{-1}(y - A\pi^{Long}) + (\pi^{Long} - \gamma)'\Omega^{-1}(\pi^{Long} - \gamma)]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}[(A\pi^{Long})'W^{-1}(A\pi^{Long}) - (A\pi^{Long})'W^{-1}y - y'W^{-1}(A\pi^{Long}) + y'W^{-1}y + (\pi^{Long})'\Omega^{-1}\pi^{Long} - (\pi^{Long})'\Omega^{-1}\gamma - \gamma'\Omega^{-1}\pi^{Long} + \gamma'\Omega^{-1}\gamma]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}[(\pi^{Long})'(A'W^{-1}A + \Omega^{-1})\pi^{Long} - (\pi^{Long})'(A'W^{-1}y + \Omega^{-1}\gamma) - (A'W^{-1}y + \Omega^{-1}\gamma)'\pi^{Long}]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}[\pi^{Long} - (A'W^{-1}A + \Omega^{-1})^{-1}(A'W^{-1}y + \Omega^{-1}\gamma)]'(A'W^{-1}A + \Omega^{-1})[\pi^{Long} - (A'W^{-1}A + \Omega^{-1})^{-1}(A'W^{-1}y + \Omega^{-1}\gamma)]\right\}$$

(for the derivation above, Assumption III guarantees W is nonsingular, so we could take inverse of it.)

Let $\gamma_N = (A'W^{-1}A + \Omega^{-1})^{-1}(A'W^{-1}y + \Omega^{-1}\gamma)$,

$$p(\pi^{Long}|y, Z) \propto \exp\left\{-\frac{1}{2}[\pi^{Long} - \gamma_N]'(A'W^{-1}A + \Omega^{-1})[\pi^{Long} - \gamma_N]\right\}$$

Thus the posterior distribution of π^{Long} is

$$\pi^{Long}|y, Z \sim N(\gamma_N, (A'W^{-1}A + \Omega^{-1})^{-1}) \tag{3.2.8}$$

Because$(A'W^{-1}A + \Omega^{-1})^{-1} = \frac{1}{N}\left(\frac{A'W^{-1}A+\Omega^{-1}}{N}\right)^{-1}$,

in which$\frac{\Omega^{-1}}{N} \to 0, and\ \frac{A'W^{-1}A}{N} \to Q$ with probability one as $N \to\infty$. By Slutsky's theorem, $(A'W^{-1}A + \Omega^{-1})^{-1} \to 0$ in the limit. Under this condition, the posterior distribution of π^{Long} is degenerated to γ_N. In general, we use γ_N as the estimate for π^{Long}, γ_N is the Bayes estimator that minimize the Bayes risk, mean square error (also called squared error risk).

$$\widehat{\pi^{Long}} = \quad \gamma_N = (A'W^{-1}A + \Omega^{-1})^{-1}(A'W^{-1}y + \Omega^{-1}\gamma) \tag{3.2.9}$$

$$= \left(\frac{A'W^{-1}A+\Omega^{-1}}{N}\right)^{-1} \left(\frac{A'W^{-1}y+\Omega^{-1}\gamma}{N}\right)$$

$$= \left(\frac{A'W^{-1}A+\Omega^{-1}}{N}\right)^{-1} \left(\frac{A'W^{-1}A\pi^{Long}+A'W^{-1}D\,\boldsymbol{\varsigma}+\Omega^{-1}\gamma}{N}\right)$$

Thus,

$$\lim_{N\to\infty} \widehat{\pi^{Long}} = Q^{-1}Q\pi^{Long} = \pi^{Long}$$

Proof:

$$\lim_{N\to\infty} \widehat{\pi^{Long}} = \lim_{N\to\infty} \left\{ \left(\frac{A'W^{-1}A + \Omega^{-1}}{N}\right)^{-1} \left(\frac{A'W^{-1}A\pi^{Long} + A'W^{-1}D\,\boldsymbol{\varsigma} + \Omega^{-1}\gamma}{N}\right) \right\}$$

$$= \left(\lim_{N\to\infty} \frac{A'W^{-1}A+\Omega^{-1}}{N}\right)^{-1} \lim_{N\to\infty} \left(\frac{A'W^{-1}A\pi^{Long}}{N} + \frac{A'W^{-1}D\,\boldsymbol{\varsigma}+\Omega^{-1}\gamma}{N}\right)$$

$$= Q^{-1}(Q\pi^{Long} + 0) = \pi^{Long}$$

Note: $\frac{A'W^{-1}D\,\boldsymbol{\varsigma}}{N} = \frac{1}{N}\sum_{i=1}^{N} \frac{d_i \otimes (1,T_i)'(1,T_i)\,\boldsymbol{\varsigma_i}}{(1,T_i)\Delta(1,T_i)'}$,

$$E\left[\frac{d_i \otimes (1,T_i)'(1,T_i)\,\boldsymbol{\varsigma_i}}{(1,T_i)\Delta(1,T_i)'}\right] = E\left[E\left(\frac{d_i \otimes (1,T_i)'(1,T_i)\,\boldsymbol{\varsigma_i}}{(1,T_i)\Delta(1,T_i)'}\middle| T_i, \boldsymbol{d}_i\right)\right] =$$

$$E\left[\frac{d_i \otimes (1,T_i)'(1,T_i)}{(1,T_i)\Delta(1,T_i)'} E\left(\boldsymbol{\varsigma_i}\middle| T_i, \boldsymbol{d}_i\right)\right] = 0$$

and the mean of posterior distribution, γ_N, is a consistent estimator of π^{Long}.

Assumption IV is strong, it gives the first and second moments of π^{Long}'s prior. For this reason, a diffuse prior is assumed.

Assumption IV': p($\boldsymbol{\pi}^{Long}$) $\propto$ const.

The posterior distribution in this case would be

$$p(\pi^{Long}|y,Z) \quad \propto \quad p(y|\pi^{Long},Z) \times p(\pi^{Long})$$

$$\propto \ \exp\{-\tfrac{1}{2}(y-A\pi^{Long})'W^{-1}(y-A\pi^{Long})\}$$

$$\propto \ \exp\left\{-\tfrac{1}{2}(A\pi^{Long})'W^{-1}(A\pi^{Long}) - (A\pi^{Long})'W^{-1}y - y'W^{-1}(A\pi^{Long})\right\}$$

$$\propto \ \exp\left\{-\tfrac{1}{2}(\pi^{Long})'A'W^{-1}A\pi^{Long} - (\pi^{Long})'A'W^{-1}y - \ (A'W^{-1}y)'\pi^{Long}\right\}$$

$$\propto \ \exp\left\{-\tfrac{1}{2}[\pi^{Long} - (A'W^{-1}A)^{-1}A'W^{-1}y]'A'W^{-1}A[\pi^{Long} - (A'W^{-1}A)^{-1}A'W^{-1}y]\right\}$$

For the derivation above, Assumption III guarantees $A'W^{-1}A$ is nonsingular.

$$\pi^{Long}|y,Z \ \sim \ N((A'W^{-1}A)^{-1}A'W^{-1}y, \ (A'W^{-1}A)^{-1}) \tag{3.2.10}$$

Again, $(A'W^{-1}A)^{-1} = \frac{1}{N}\left(\frac{A'W^{-1}A}{N}\right)^{-1} \to 0, \ as \ N \ \to \infty$, the posterior distribution for π^{Long} degenerates to its mean $(A'W^{-1}A)^{-1}A'W^{-1}y$, which converges to π^{Long}. In this diffuse prior case, we say

$$\widehat{\pi^{Long}} = (A'W^{-1}A)^{-1}A'W^{-1}y \tag{3.2.11}$$

It is straightforward to show that (3.2.11) provides us a consistent estimator for π^{Long} following same steps as before and, moreover, it requires a relatively weaker assumption on the prior. By the central limit theorem we could derive the asymptotic distribution of $\widehat{\pi^{Long}}$ in (2.11),

$$N^{0.5}(\widehat{\pi^{Long}} - \pi^{Long}) \sim N(0, (A'W^{-1}A)^{-1}) \quad (3.2.12)$$

When a diffuse prior is used, our Bayes estimator is just the maximum likelihood (MLE) estimator, which would asymptotically achieve the Cramer-Rao lower bound, and thus our estimator $\widehat{\pi^{Long}} = (A'W^{-1}A)^{-1}A'W^{-1}y$ is efficient as well. Once the estimate for π^{Long} is known, we can substitute it back into (3.2.3) to generate an estimate for $\boldsymbol{\beta}_i$, the CVC estimator, whose second element is just the treatment effect for individual i. This estimator is also a very familiar estimator to statisticians (see Swamy and Mehta 1975; Swamy and Tinsley 1980). All the unknown quantities of model (3.2.5) including Δ are estimated by iteratively evaluating $D\hat{\varsigma}$, Δ based on $D\hat{\varsigma}$, $\widehat{\pi^{Long}}$, and $\hat{\beta}$.

One additional point worth mentioning about formula (2.9) and (2.11) is that the Bayes estimator (either (3.2.9) or (3.2.11)) for $\boldsymbol{\pi}^{Long}$ is the mean of posterior distribution and is also the solution to the argmin of $\mathrm{E}_{\pi^{\mathrm{Long}}}\{\mathrm{E}_{y|\pi^{\mathrm{Long}}}[(\widehat{\pi^{\mathrm{Long}}} - \pi^{\mathrm{Long}})(\widehat{\pi^{\mathrm{Long}}} - \pi^{\mathrm{Long}})']\}$, see Lehmann and Casella (1998, theorem 4.1.1). Swamy and Tinsley (1980) also pointed out that this estimator for $\boldsymbol{\pi}^{Long}$ minimized the same quadratic form from the BLUP perspective.

3.2.4. Estimation

An alternative to the above approach is to update W, which is just $D\hat{\varsigma}\ D\hat{\varsigma}'$. The problem is that W is a N × N matrix and it imposes too much of a computational workload when the sample size N is large, especially in the Monte Carlo simulations provided in section 3. Moreover, $D\hat{\varsigma}$ as error term should be relatively small, which may

result in $D\hat{\varsigma}\, D\hat{\varsigma}'$ having some relatively small nonzero eigenvalues, so W would be very close to a singular matrix in computation. As a result, the calculation of $\widehat{\pi^{Long}}$, which requires inverse of W, is infeasible.

Δ is a good option for updating, in fact updating Δ_{11} and $(\Delta_{12}+\Delta_{21}+\Delta_{22})$ is enough. $W = \mathrm{D}(I_N \otimes \Delta)D'$ is a diagonal matrix by derivation, with i-th diagonal element being $\Delta_{11} + (\Delta_{12}+\Delta_{21})T_i + \Delta_{22}T_i^{\,2}$. When T_i=0, it is Δ_{11}, while for T_i =1, it is $\Delta_{11}+\Delta_{12}+\Delta_{21}+\Delta_{22}$. The stability of Δ_{11} and $(\Delta_{12}+\Delta_{21}+\Delta_{22})$ freeze W. So we will keep updating Δ_{11} and $(\Delta_{12}+\Delta_{21}+\Delta_{22})$ until the difference of them between two successive iteration is small enough. We start the iteration with the initial value $\Delta = I_2$, an identity matrix of order 2. In each iteration, we set $\Delta_{11} = \frac{\sum_{i:\,T_i=0}(D\hat{\varsigma})_i^2}{\sum_{i:\,T_i=0}1}$ and $\Delta_{11}+\Delta_{12}+\Delta_{21}+\Delta_{22} = \frac{\sum_{i:\,T_i=1}(D\hat{\varsigma})_i^2}{\sum_{i:\,T_i=1}1}$. As for the criterion that evaluates the difference, we set it equal std(y)$\times 10^{-10}$. Once the differences of both Δ_{11} and $(\Delta_{12}+\Delta_{21}+\Delta_{22})$ between the two successive iterations are smaller than my criterion, we stop iteration, and use the latest of $\widehat{\pi^{Long}}$ and (3.2.3) to get estimates for $\hat{\boldsymbol{\beta}}_i$.

The average impact of the treatment on an individual is given by the average treatment effect (ATE), averaged over the entire sample:

$$\text{ATE} = E(y_{i1} - y_{i0}) \tag{3.2.12}$$

An estimate of ATE from the equation $\hat{\boldsymbol{\beta}}_i = \hat{\Pi}\boldsymbol{d}_i + \hat{\boldsymbol{\varsigma}}_i$ is approximated by $\frac{\sum_i \widehat{\beta_{11}}}{N}$.

The average treatment effect on the treated (ATT) is

$$\text{ATT} = E[(y_{i1} - y_{i0}) \mid T_i = 1] \tag{3.2.13}$$

In the case where the treatment is not completely randomly assigned, ATE ≠ ATT. It is not possible to calculate ATT if assignment to the treatment is endogenous. With model (3.2.5) this difficulty does not arise. An estimate of ATT from $\hat{\boldsymbol{\beta}}_i = \hat{\Pi}\boldsymbol{d}_i + \hat{\boldsymbol{\varsigma}}_i$ is $\frac{\sum_{i:T_i=1} \widehat{\beta_{i1}}}{\sum_{i:T_i=1} 1}$.

3.3. Simulation Experiment

In this section, we generate nonexperimental datasets using three different functional forms of h_0 and h_1, with sample size N=100, 500, 1000, and tried four different sets of coefficient drivers in estimation.

3.3.1. Assumed Data Generating Process

The models considered for our experiments were the following: X = (X_1, X_2, X_3, X_4, X_5)′ ; (X_1, X_2, X_3)′ is multivariate normal with mean vector (0, 0, 0)′ and covariance matrix $\begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{pmatrix}$; X_4 is uniform over the interval (-3, 3); X_5 is chi-square with one degree of freedom, denoted by $\chi^2[1]$; X_4 and X_5 are independent of (X_1, X_2, X_3)′.

The controlled (or untreated) and treated groups of individuals were determined by assigning each observation according to

$$T = 1(X_1 + 2X_2 - 2X_3 - X_4 - 0.5X_5 + \varepsilon > 0) \qquad (3.3.1)$$

this equation means that $T = 1$ if the condition within parentheses is true and T = 0 otherwise. For the same X_i, the distribution of error term ε determines the number of treated and untreated observations, and four different error distributions for ε were considered:

(i) N(0,10); (ii) N(0,100); (iii) $\chi^2[5]$; (iv) 50-50 mixture of $N(-5,10)$ and $N(5,10)$

(In simulation, the four distributions of ϵ_i make the number of treated units account for 42.8%, 49.2%, 69.4% and 41.8% of the sample respectively)

As in Section 2.1, we considered

$$Y = TY_1 + (1-T)Y_0$$

$$Y_0 = h_0(X) + \upsilon_0 \qquad v_0 \sim N(0,1)$$

$$Y_1 = h_1(X) + \upsilon_1 \qquad v_1 \sim N(0,1)$$

The unit variances for υ_0 and υ_1 provide us with a scale normalization that simplifies our calculations with simulated data. We considered three different functional forms of h_0 and h_1.

Both h_0 and h_1 are linear:

$$h_0(X) = 3X_1 + X_2 + 2X_3$$

$$h_1(X) = 2X_1 + 5X_2 + 3X_3$$

Both h_0 and h_1 are quadratic functions of X:

$$h_0(X) = 3X_1 + X_2^2 + 2X_3$$

$$h_1(X) = 2X_1 + 5X_2^2 + 3X_3$$

Both h_0 and h_1 are nonlinear:

$$h_0(X) = \exp\{X_1\} + X_2^2 - X_3$$

$$h_1(X) = X_1^2 - X_1 \times X_2 + 3X_3 - \log(X_5)$$

Choice of specific forms and numbers of coefficient drivers when analyzing real-world data is of course an unresolved issue. We experiment with various sets of possible coefficient drivers and compare their results before we settle on one set. In this connection, we emphasize the role of simulation experiments such as this one in giving us ideas about what methods of selecting the coefficient drivers can be followed in the real-world settings.

Four different vectors of coefficient drivers were considered.

$$\boldsymbol{d}_1 = (X_1, X_2, X_3, X_4, X_5)'$$

$$\boldsymbol{d}_2 = (X_1, X_2^2, X_3, X_4, X_5)'$$

$$\boldsymbol{d}_3 = (X_1^2, X_2^2, X_3, X_4^2, X_5^2)'$$

$$\boldsymbol{d}_4 = (X_k, X_k \times X_j)' \text{ where } k, j = 1, \ldots, 5$$

The vector $\boldsymbol{d}_1$ is the same as the vector X and hence is the vector of most appropriate coefficient drivers for the linear h_i case. In this case, the vectors $\boldsymbol{d}_2$ and $\boldsymbol{d}_3$ are inappropriate. For the quadratic h_i case, $\boldsymbol{d}_2$ is appropriate and $\boldsymbol{d}_1$ and $\boldsymbol{d}_3$ are inappropriate. For the nonlinear h_i case, $\boldsymbol{d}_3$ is closer to the vector of most appropriate coefficient drivers than $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$. The vector $\boldsymbol{d}_4$ is comprehensive in the sense that it has $\boldsymbol{d}_1$ through $\boldsymbol{d}_3$ as its sub vectors. Good sets of coefficient drivers should satisfy assumption I and II, and gives good approximation. In the simulation, the treatment effects as well as *y* is known, so we will check the goodness of fit from both perspectives.

With four distributions of ε in (3.3.1), three functional forms of the pair ($h_0(X)$, $h_1(X)$), and four vectors of coefficient drivers, we have 48 distributions. From each of these distributions 1000 samples of sizes N = 100, 500, and 1000 were drawn randomly. An observation vector in each sample refers to an individual. For example, a sample of size 100 means that observation vectors for 100 individuals are available. In our experiments, we had observations on all the variables in $(Y, Y_0, Y_1, T, X)'$ but used only the observations on the variables in $(Y, T, X)'$ to replicate the real situations where the observations on both Y_0 and Y_1 for the same individual are not available.

3.3.2. Goodness of Fit

In this section, we address the question: How close is the estimated model: $\hat{\boldsymbol{y}} = A\,\hat{\boldsymbol{\pi}}^{Long} + \mathrm{D}\hat{\boldsymbol{\varsigma}}$ with all of its underlying assumptions to model (3.2.1)? To answer this

question, we use coefficient of determination R^2 to gauge the fitness. The values of R^2 reported in Table 1 are obtained by averaging R^2 of the 1000 simulations.

It can be seen from Table 3.1 that

(i) The comprehensive vector of coefficient drivers d_4 produces the highest R2 regardless of the sample size and ε-distribution in every linear, quadratic, or nonlinear h_i case we considered.

(ii) If we consider only d_1, d_2, and d_3, the vector d_1 of appropriate coefficient drivers in the linear h_i cases, the vector d_2 of appropriate coefficient drivers in the quadratic h_i cases and the vector d_3 of coefficient drivers in the nonlinear h_i cases produce the highest R2 regardless of the sample size and ε-distribution.

(iii) The comprehensive vector d_4 of coefficient drivers produces the R2 s that are bigger than those produced by the appropriate vector d_1 in the linear h_i cases, by the appropriate vector d_2 in the quadratic h_i cases and by d_3 in the nonlinear h_i cases.

The practical guideline, the values in Table 3.1, suggest to choose appropriate coefficient drivers when possible, otherwise comprehensive vectors of coefficient drivers would yield relatively better fit.

The closeness of estimated and true treatment effects is also checked. Inserting (3.2.10) into (3.2.3) gives $\hat{\boldsymbol{\beta}}_i = \hat{\Pi}\boldsymbol{d}_i + \hat{\boldsymbol{\varsigma}}_i$. Note that this estimate is obtained because of

the unique capability of the CVC model in (3.2.4). This capability is that model (3.2.4) can estimate β_{i1}, the treatment effect, for every individual under study, even in the absence of data on both y_{i0} and y_{i1} for each individual.

Adjusted	N=100				N=500				N=1000			
R2	N(0,1)	N(0,10)	$\chi^2[5]$	bimodal	N(0,1)	N(0,10)	$\chi^2[5]$	bimodal	N(0,1)	N(0,10)	$\chi^2[5]$	bimodal
Linear												
CVC $\boldsymbol{d}_1$	0.9598	0.9611	0.9661	0.9592	0.9592	0.9617	0.9654	0.9592	0.9593	0.9617	0.9651	0.9593
CVC $\boldsymbol{d}_2$	0.7736	0.7415	0.7417	0.7926	0.7637	0.7157	0.7367	0.7737	0.7609	0.7184	0.7322	0.7734
CVC $\boldsymbol{d}_3$	0.2948	0.1894	0.4759	0.3551	0.2563	0.1074	0.4533	0.2919	0.2402	0.0986	0.4485	0.2925
CVC $\boldsymbol{d}_4$	0.9613	0.9625	0.9678	0.9622	0.9597	0.9616	0.9655	0.9593	0.9597	0.9616	0.9656	0.9597
quadratic												
CVC $\boldsymbol{d}_1$	0.4984	0.3289	0.5158	0.5511	0.4455	0.2522	0.4837	0.4943	0.4402	0.2416	0.4923	0.4815
CVC $\boldsymbol{d}_2$	0.9690	0.9721	0.9743	0.9720	0.9728	0.9734	0.9759	0.9731	0.9730	0.9739	0.9768	0.9734
CVC $\boldsymbol{d}_3$	0.8102	0.8023	0.9123	0.8029	0.8094	0.7991	0.9143	0.8137	0.8111	0.8001	0.9162	0.8117
CVC $\boldsymbol{d}_4$	0.9731	0.9721	0.9748	0.9742	0.9729	0.9733	0.9761	0.9727	0.9732	0.9741	0.9767	0.9738
nonlinear												
CVC $\boldsymbol{d}_1$	0.3942	0.4507	0.3864	0.3721	0.3131	0.3858	0.3611	0.3361	0.3124	0.3768	0.3465	0.3158
CVC $\boldsymbol{d}_2$	0.5863	0.6509	0.4475	0.5936	0.5488	0.6085	0.4234	0.5470	0.5423	0.5836	0.4261	0.5373
CVC $\boldsymbol{d}_3$	0.6762	0.7602	0.5975	0.6441	0.6266	0.7213	0.5873	0.6115	0.6123	0.6971	0.5875	0.6011
CVC $\boldsymbol{d}_4$	0.8548	0.8848	0.7489	0.8511	0.8189	0.8473	0.7453	0.8136	0.8072	0.8228	0.7425	0.8030

Table 3.1 Adjusted R^2

To make the comparison of true and estimated treatment effects straightforward, we can examine such diagnostics as the scatter plot of true treatment effect versus estimated treatment effect in Figure 1, which is for the case of nonlinear $h_0(X)$ and $h_1(X)$ functions, the comprehensive coefficient driver vector $\boldsymbol{d}_4$, and the sample size 1000 (it corresponds to the last four cells of the bottom line in Table 1), since these sample characteristics would appear to be more like what we would encounter in a real empirical setting. The upper left plot is for N(0, 10) error term, upper right is for N(0,100), lower left is for $\chi^2[5]$, and lower right is for bimodal error term. The scatter plots are concentrated around 45 degree lines, which means estimated treatment effect is close to the true value. Even though the fits for majority points are good, the method does not provide good estimates for the boundary points of treatment effects, see the tails in lower left plot. This happens for the case in which the comprehensive coefficient drivers set $\boldsymbol{d}_4$, for nonlinear $h_0(X)$ and $h_1(X)$ functions, is used. If we use appropriate coefficient drivers, say $\boldsymbol{d}_1$ in the linear h_i cases, the vector $\boldsymbol{d}_2$ in the quadratic h_i cases, we do not have this kind of boundary problem and all points are very close to the 45 degree line. As mentioned in section 3.3.1, it is quite difficult to find the most appropriate coefficient drivers in practice, and thus we might have to use the coefficient driver set that gives the relatively higher R^2.

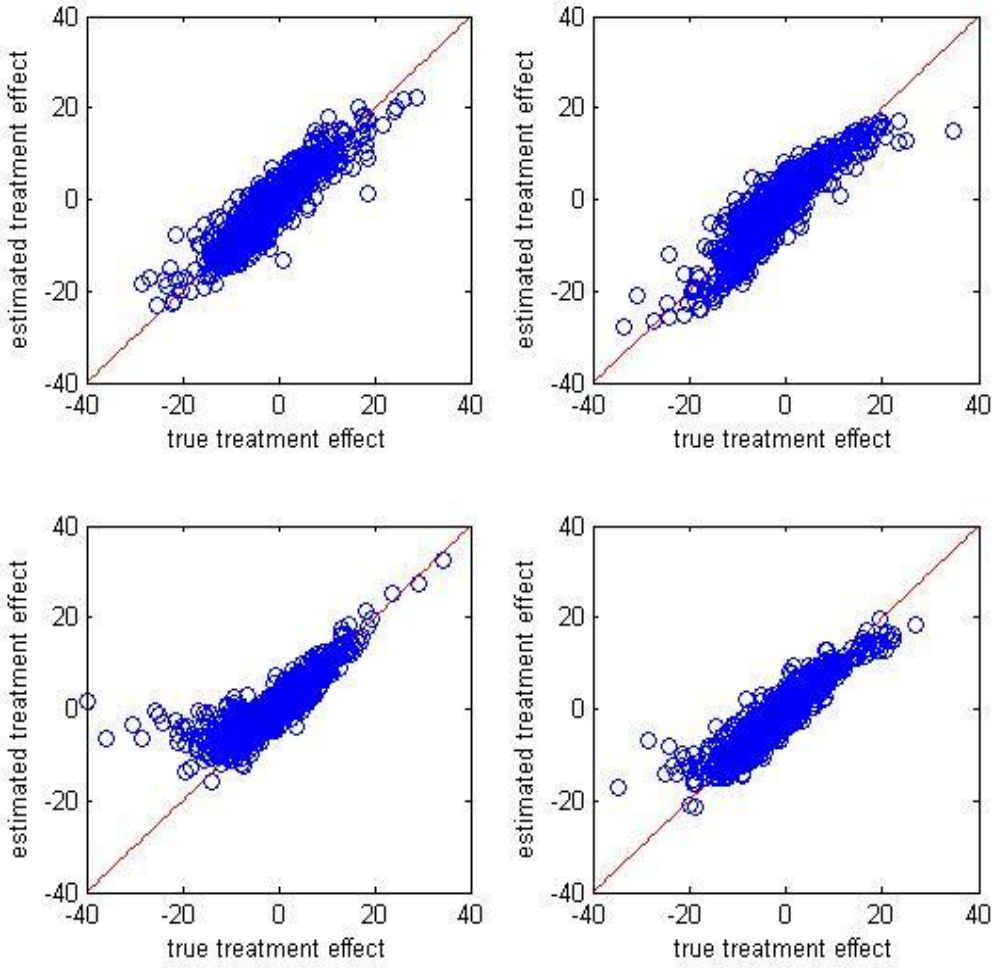

Figure 3.1 Estimated vs true treatment effects

(N = 1000, nonlinear h_i's, d_4)

For a variety of reasons program evaluators may be more interested in the distribution of the treatment effect rather than individual values since the distribution provides an inclusive measure of the percentage of people who benefit from the program. The quantile regression method gives the treatment effect on the quantiles of the outcome rather than the treatment effect distribution. One might propose a deconvolution method

to estimate the treatment effect distribution. However, the deconvolution method is only reliable when the treatment is randomly assigned, while in our case of self-selection into treatment, the deconvolution method fails because this condition of random assignment is not met. Figures 3.2-3.5 display kernel density functions of estimated treatment effect $\hat{\beta}_{i1}$ (= estimated β_{i1}) estimated by our proposed method and the histograms of the true treatment effect β_{i1} for the general case of nonlinear $h_0(X)$ and $h_1(X)$ functions, the comprehensive coefficient driver vector $\boldsymbol{d}_4$, and the sample size 1000. We can see that the estimated density tracks the data very closely.

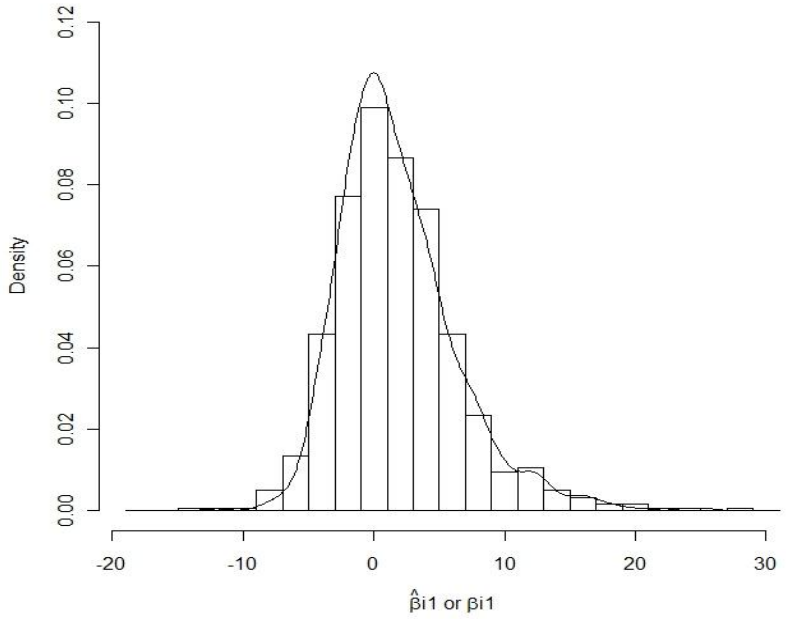

Figure 3.2 Kernel density for $\hat{\beta}_{i1}$ (N = 1000, nonlinear h_i's, $\boldsymbol{d}_4$, N(0,10))

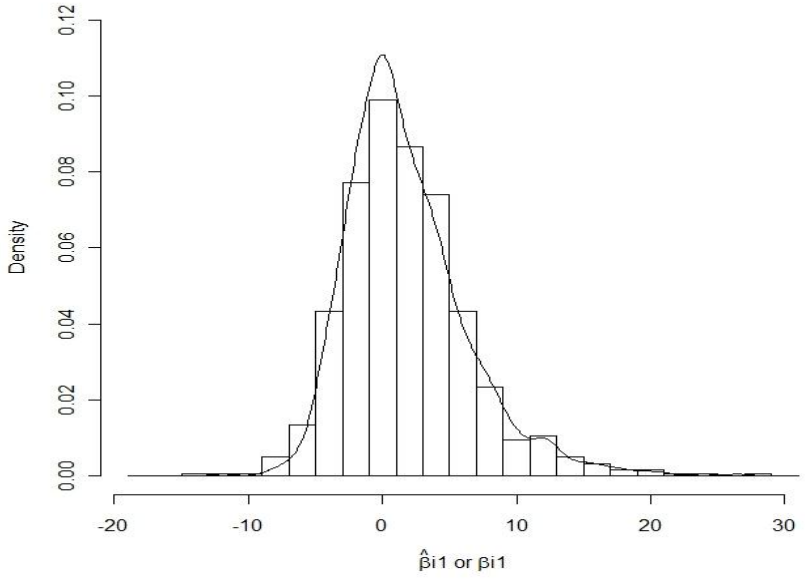

Figure 3.3 Kernel density for $\hat{\beta}_{i1}$ (N = 1000, nonlinear h_i's, $\boldsymbol{d}_4$, N(0,100))

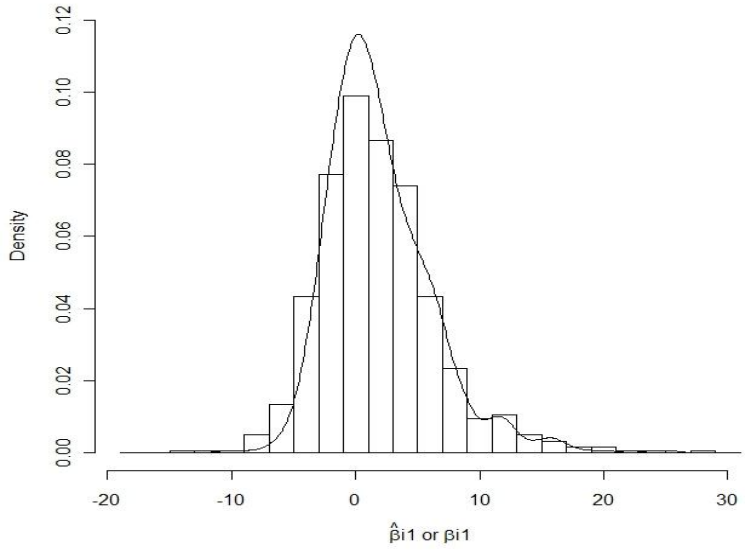

Figure 3.4 Kernel density for $\hat{\beta}_{i1}$ (N = 1000, nonlinear h_i's, $\boldsymbol{d}_4$, $\chi^2[5]$)

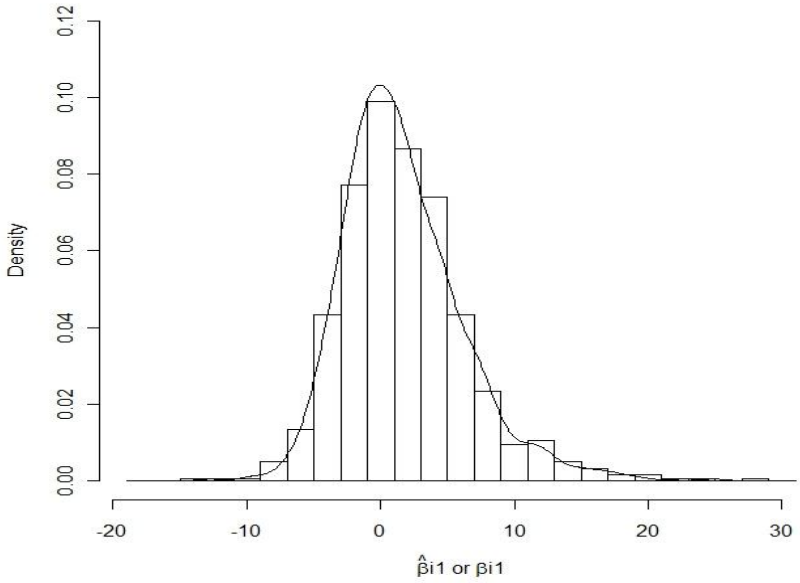

Figure 3.5 Kernel density for $\hat{\beta}_{i1}$ (N = 1000, nonlinear h_i's, d_4,bimodal)

3.3.3. Comparison with Matching Estimator in ATT and ATE Estimations

The average impact of the treatment on an individual drawn at random from the entire population is given by the average treatment effect (ATE), averaged over the entire population:

$$\text{ATE} = E(y_{i1} - y_{i0}) \tag{3.3.1}$$

An estimate of ATE from the equation $\hat{\boldsymbol{\beta}}_i = \hat{\Pi}\boldsymbol{d}_i + \hat{\boldsymbol{\varsigma}}_i$ is

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_{i1} \tag{3.3.2}$$

The average treatment effect on the treated (ATT) is

$$\text{ATT} = E[(y_{i1} - y_{i0}) \mid T_i = 1] \tag{3.3.3}$$

The difficulty of measuring this without model (3.2.5) is that the counterfactual $E[y_{i0} \mid T_i = 1]$ is unknown. In the case where the treatment is not completely randomly assigned, ATE ≠ ATT. It is not possible to calculate ATT if assignment to the treatment is endogenous. With model (3.2.5) this difficulty does not arise. An estimate of ATT from $\hat{\boldsymbol{\beta}}_i = \hat{\Pi}\boldsymbol{d}_i + \hat{\boldsymbol{\zeta}}_i$ is

$$\frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_{i1} \tag{3..3.4}$$

where only the estimates $\hat{\beta}_{i1}$ corresponding to $T_i = 1$ are used and M is the number of such $\hat{\beta}_{i1}$. we will compare the estimates of (3.3.2) and (3.3.4) with estimates of matching method in the following part.

Rosenbaum and Rubin (1983) suggested the use of propensity score matching to estimate ATE and ATT. Propensity score matching (PSM): $\boldsymbol{p}(\boldsymbol{x}_i) = \Pr(\boldsymbol{T}_i = 1 \mid \boldsymbol{x}_i) = \boldsymbol{E}(\boldsymbol{T}_i \mid \boldsymbol{x}_i)$ is the probability that individual i is treated. Individuals with similar propensity scores are paired to compute the ATE and ATT. The PSM uses the information contained in $\boldsymbol{x}_i$ in the form of $\boldsymbol{p}(\boldsymbol{x}_i)$ to pair individuals.

Assumption V: $(y_{i1}, y_{i0}) \perp T_i \mid p(x_i)$.

Under this assumption,

$$\text{ATE} = E_{p(X)}\{E[Y_1 \mid T=1, p(X)] - E[Y_0 \mid T=0, p(X)]\} \tag{3.3.5}$$

$$\text{ATT} = E_{p(X)|T=1}\{E[Y_1 \mid T=1, p(X)] - E[Y_0 \mid T=0, p(X)]\} \tag{3.3.6}$$

From (3.3.2) and (3.3.4), it is obtained that

$$\widehat{ATE} = \frac{N_T}{N}\widehat{ATT} + \frac{N_C}{N}\widehat{ATC} \tag{3.3.7}$$

$$\widehat{ATT} = \frac{1}{N_T}\sum_{i\in T}\ [\ y_{i1} - \sum_{j\in C(i)}\ W_{ij}\ y_{i0}\ ] \tag{3.3.8}$$

$$\widehat{ATC} = \frac{1}{N_C}\sum_{i\in C}\ [\sum_{j\in T(i)}\ W_{ij}\ y_{i1} - \ y_{i0}] \tag{3.3.9}$$

where C stands for the set of control units, i.e., untreated units; N_C is the number of units in set C; similarly, for T and N_T; $C(i)$ is the set of control units matched to the treated unit i with an estimated value of the propensity score $\boldsymbol{p}(\boldsymbol{x_i})$ and $N_{C(i)}$ denotes the number of units in $C(i)$; similarly for $T(i)$ and $N_{T(i)}$; $W_{ij} = \frac{1}{N_{C(i)}}$ if $i \in T$ and $= \frac{1}{N_{T(i)}}$ otherwise; $C(i) = \{j: |p(x_i) - p(x_j)| \leq |p(x_i) - p(x_k)|\}$ where k is the m-th nearest matched unit}; similarly for $T(i)$.

In addition to PSM, we also consider Mahalanobis matching. In Mahalanobis matching the set of control units matched to the treated unit i is given by

$C(i)$ = {: $D_{ij} \leq D_{ik}$, where k is the m-th nearest matched unit}, where D_{ij} = $(\boldsymbol{x_i} - \boldsymbol{x_j})'\boldsymbol{S}^{-1}(\boldsymbol{x_i} - \boldsymbol{x_j})$, and where $\boldsymbol{S}$ is the sample covariance matrix of the matching

variables from the set (C or T) that does not contain i and similarly for $T(i)$. All the other definitions are common to the PSM and Mahalanobis matching.

The estimates of ATE and ATT based on the PSM and Mahalanobis matching are reported in Table 2 for the cases involving $\boldsymbol{d}_4$ and N = 1000. All the estimates of ATE (or ATT) in Table 2 are expressed in relative terms; that is, relative to the corresponding true values of ATE (or ATT), which serve as our primary benchmark. For this reason, the true values of ATE and ATT in Table 2 are 1 for all the cases we considered. True values of (3.3.2) and (3.3.4) are derived from the parameterization used in our data generating process. It can be seen from Table 3. 2 that in 42 of 48 cases, the estimates of ATE and ATT yielded by the CVC $\boldsymbol{d}_4$ model in (3.2.4) are closer (sometimes much closer) to the true value 1 than those yielded by the PSM or the Mahalanobis matching. The estimates of ATE and ATT given by the various methods are plotted in Figure 6-8, which clearly demonstrate the potential advantages of the CVC $\boldsymbol{d}_4$ model over other matching approaches in analyzing treatment effects we consider herein.

Table 3.2: Estimates of ATE and ATT (N = 1000)

	ATE				ATT			
	N(0,10)	N(0,100)	$\chi^2[5]$	bimodal	N(0,10)	N(0,100)	$\chi^2[5]$	bimodal
linear								
TRUE	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
PSM	1.6434	-0.7734	-5.7295	0.3824	1.1126	-57.1298	0.1540	1.3080
Mahalanobis	1.1727	0.6262	-11.9510	4.9305	1.3055	1.1023	3.8643	1.3026
CVC $\boldsymbol{d}_4$	1.0872	0.9184	-0.2441	1.0164	1.0163	0.6152	0.9483	1.0093
quadratic								
TRUE	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
PSM	0.9896	0.9923	1.1750	0.9634	1.0360	1.0861	0.9760	1.0432
Mahalanobis	0.9031	0.9401	1.1002	0.8867	1.0766	1.0347	1.5613	1.0838
CVC $\boldsymbol{d}_4$	1.0008	1.0000	0.9946	0.9996	1.0000	1.0001	0.9871	0.9998
nonlinear								
TRUE	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
PSM	1.0053	0.994816	1.3646	0.9622	1.0606	1.1374	38.1328	1.0662
Mahalanobis	0.8983	0.9803	1.3595	0.8547	1.4068	1.1198	16.2953	1.7564
CVC $\boldsymbol{d}_4$	0.9605	0.9613	1.2715	0.9580	0.7453	0.9583	7.8429	0.8212

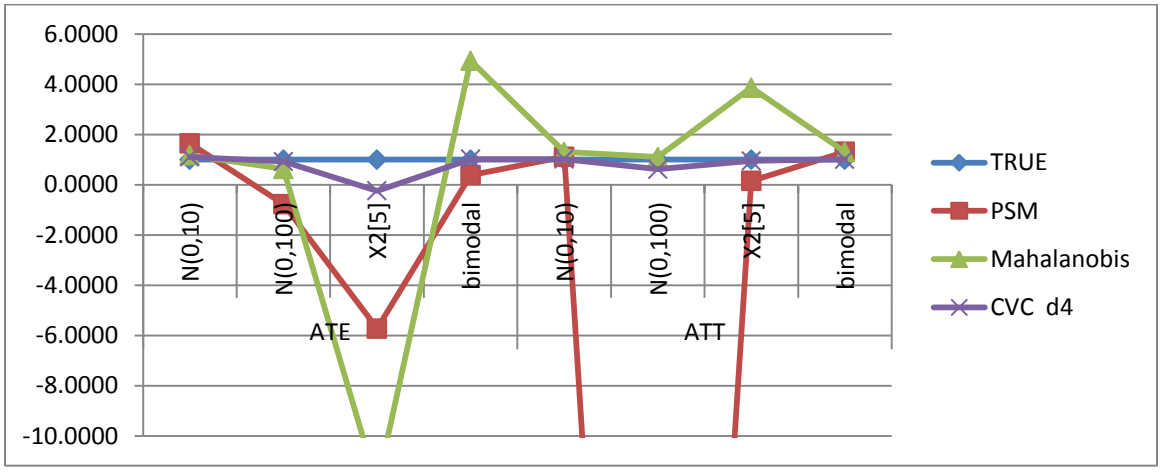

Figure 3.6 ATE and ATT (linear h_i 's)

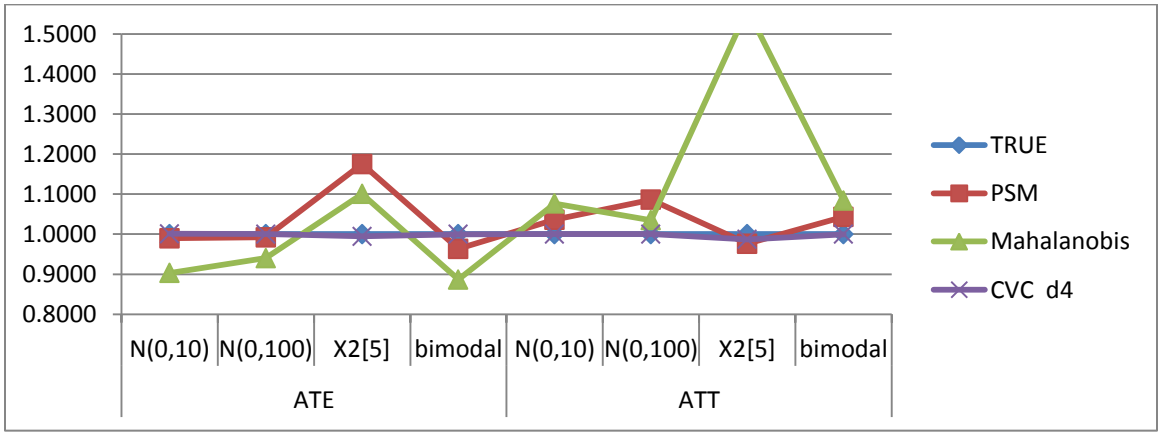

Figure 3.7 ATE and ATT (quadratic h_i 's)

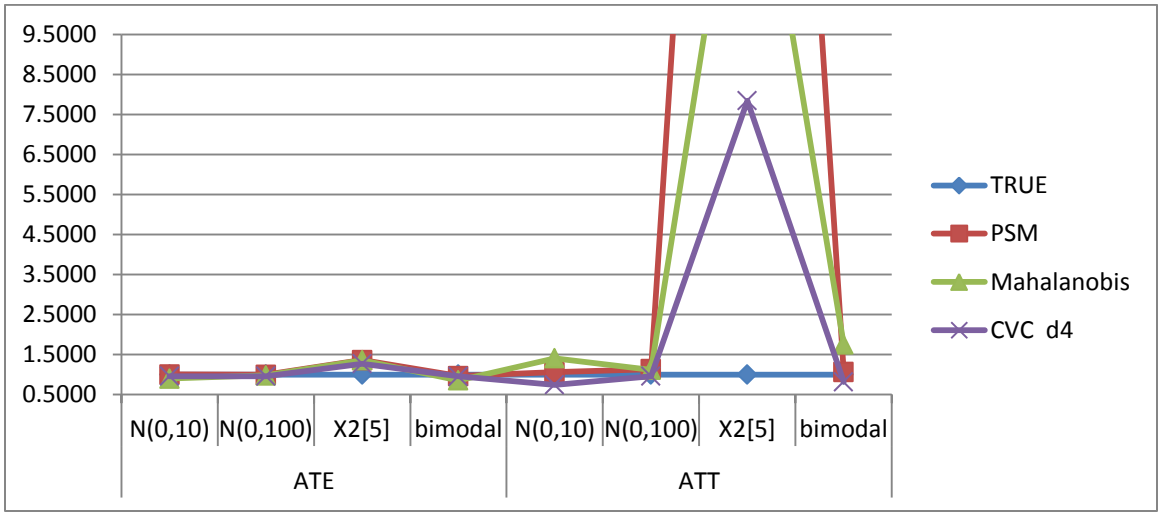

Figure 3.8 ATE and ATT (nonlinear h_i's)

3.4. Empirical Application

In this section, we use the proposed estimator to examine the impact distribution on two datasets. The first concerns the National Job Training Partnership Act (JTPA) Study. The National JTPA study is financed by U.S. Department of Labor to evaluate the effectiveness of training programs funded under Title II-A of the Job Training Partnership Act of 1982. The training program included classroom training, on-the-job training and job search assistance to the disadvantaged. The second dataset is on firms' seasoned equity offering (SEO), and checks the SEO decision's impact on firms' operational performance, which is reflected in the operating income before depreciation and amortization over asset (OIBD/asset).

3.4.1. Application With National JTPA Data

The experimental sample of National JTPA includes 20,601 applicants to JTPA in the 16 study JTPA training centers who were accepted by JTPA staff and randomly assigned to a treatment group, which was allowed to enter the program, and a control group, which was excluded from the JTPA program for 18 months. This is a simple randomized case for treatment effect analysis. Follow-up interviews were conducted with each person in the experimental sample during the period from 12-24 months after random assignment. The interview collected information on employment, earnings, participation in government transfer programs, as well as information on schooling and training during the period after random assignment. Following Heckman et al. (1997), we use observations on adult women (women aged 22 or more), and present the impacts of training on their self-reported earnings in the eighteen months following random assignment.

In this scenario, the treatment is the training program, and the treated group consists of adult women who attended the program, while the control group includes the ones that did not attend. The variable of interest is the self-reported earnings in the eighteen months following random assignment. Age, sex, race, educational level, marital status, number of dependent children, earnings in the past year, weeks worked in the past year, wage at most recent job, hours worked per week at most recent job and welfare history are observable variables, conditional on which could earnings in 18 months and treatment be independent, as claimed by the unconfoundedness assumption. Our sample consists of 2541 observations for adult women. The set of coefficient drivers is

constructed by including these variables, their squared terms and cross terms step by step, each step adding one more term. After including the variables themselves, their squared terms and the cross terms of education, marital status, number of dependent children, weeks worked in the past year, wage at most recent job, hours worked per week at most recent job and welfare history as the coefficient drivers, the distribution becomes stable, adding new coefficient drivers does not change the distribution and R^2 in any discernible way, we refer to this set of coefficient variables as set (1). The set (2) is set (1) plus age times every other term while set (3) is set (2) and the remaining the cross terms. The properties of the impact distribution using these three different sets of coefficient drivers are listed in Table 3.3.

Table 3.3 Properties of impact distribution with different coefficient drivers (National JTPA study 18 month impact sample: adult women)

	min	median	mean	max	Standard deviation	Percent positive	Adjusted R^2
Set (1)	-21.4108	0.3444	0.4035	15.6956	1.7103	0.6112	0.6436
Set (2)	-23.2744	0.3400	0.3849	15.9617	1.8245	0.6076	0.6443
Set (3)	-22.4542	0.4082	0.4085	15.2037	2.1162	0.6009	0.6545

From minimum, maximum, mean median and etc., we can see that the impact distributions using coefficient driver sets (1), (2) and (3) are very close, and including extra terms only marginally increases R^2. Moreover, the correlations of the estimated

error terms and the coefficient drivers are very small, they are significantly uncorrelated, and thus this impact distribution would appear to be valid. Because we have so many coefficient drivers (for set (3) we have 67 coefficient drivers), the statistics and plots for correlation test are not provided to save space but are available upon request. From the descriptive properties of distribution, it is relatively straightforward to see that the impact distribution has positive mean and median and that more than 60 percent of population could increase their earnings through the training of the program. Heckman et al. (1997) gives very close estimates to this method for the impact distribution, while my method additionally gives estimates of treatment effect on each individual.

Figure 3.9 displays the smoothed estimated distribution of the treatment effect with coefficient driver set (3), while the cumulative distribution function is shown in Figure 3.10. From Figure 3.9 we can see that more than half of the impact is positive indicating that more than half of the population benefited from the training program. Figure 3.10 shows the cumulative distribution function of the impact compared with normal distribution's c.d.f.. From these two figures, we see that the impact distribution is more concentrated around zero than the normal distribution with same mean and variance.

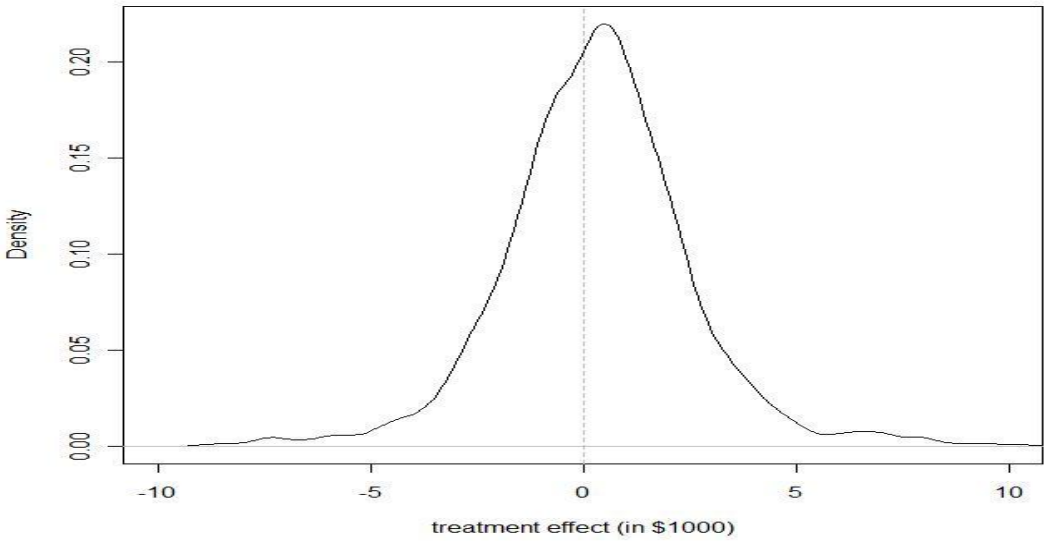

Figure 3.9 smoothed estimated impact density

(National JTPA study 18 month impact sample: adult women)

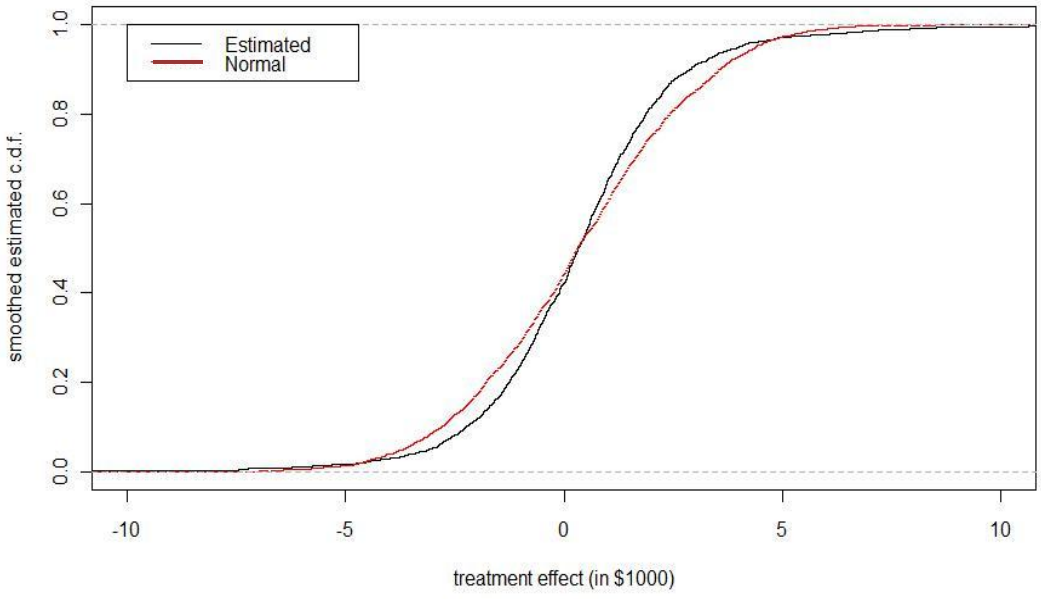

Figure 3.10 smoothed estimated c.d.f of impact and normal c.d.f.

(National JTPA study 18 month impact sample: adult women)

Table 3.4 gives estimates of the impact distribution's percentile, mean, the distribution's standard deviation, and positive percentages using 10000 bootstrapped samples. The numbers in parenthesis are standard deviations based on the bootstrap. The bootstrap results are very close to the results in Table 3.3 and support the conclusions above. Moreover, the bootstrap errors for 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} are all very small implying that the estimate of the impact distribution is tight, although its minimum and maximum values could change substantially.

Table 3.4 Bootstrap results for impact distribution
(National JTPA study 18 month impact sample: adult women)

5^{th} percentile	25^{th} percentile	50^{th} percentile	7 5^{th} percentile	95^{th} percentile
-3.829	-1.076	0.361	1.770	4.705
(0.571)	(0.255)	(0.208)	(0.237)	(0.547)
Minimum	Maximum	Mean	Dist. Std dev	Percent pos.
-27.766	25.755	0.383	3.046	0.570
(15.676)	(10.922)	(0.233)	(0.373)	(0.040)

One more interesting point is that densities of the treatment effects are far different when we plot them by races as can be seen in Figure 3.11. The impact distribution of White and Black are very close to the density of whole sample, but Hispanic is left positive skewed while OTHER RACE is negative skewed. In fact, only

22.27% of Hispanics receive a positive treatment effect while as much as 94.55% of the other races gain from the program.

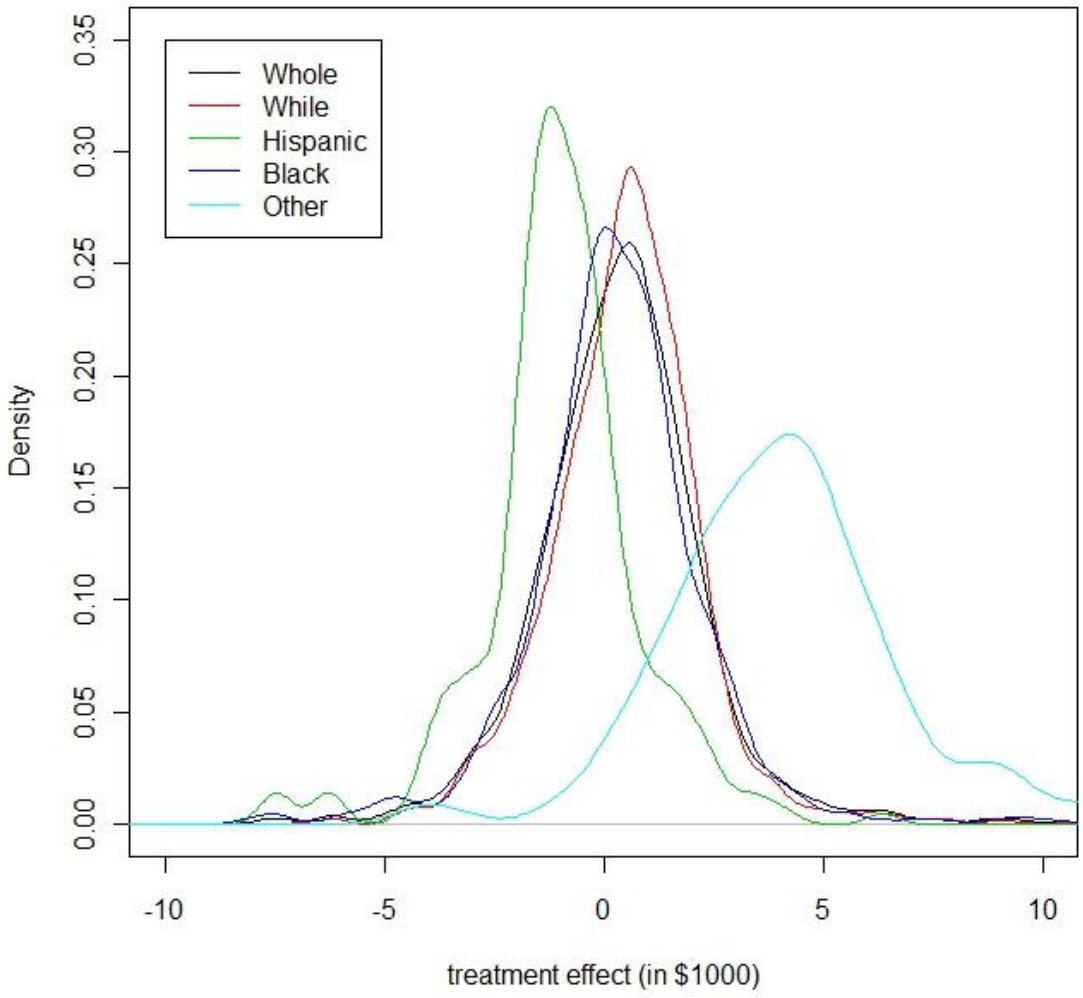

Figure 3.11 Impact distributions by race

(National JTPA study 18 month impact sample: adult women)

3.4.2. Application With SEO Data

In this section, we use the proposed estimator to examine the impact distribution of the operating income before depreciation and amortization over asset (OIBD/asset) by a firm's choice of seasoned equity offerings. Loughran and Ritter (1997) used 1338 seasoned equity offerings during 1979-1989, and reveal that firms who issue seasoned equities tend to have relatively low OIBD/asset than their counterparts who did not issue a seasoned equity offering (SEO). SEO's are an indicator of poor subsequent operating performance, including OIBD/asset. we use data from 1999-2009 with 1077 seasoned equity offerings together with the proposed method to examine the impact of SEO on ROA, and check whether the majority firm's OIBD/asset deteriorates after an SEO.

3.4.2.1. Data Selection Rule

The SEO observations during 1999 through 2009 are collected from the Securities Data Corporation (SDC) New Issues database. They must meet the following criteria to be used in my analysis.

(1) We only include securities identified by Center for Research in Securities Prices (CRSP) as ordinary common stock (share codes 10 and 11).

(2) The company is listed on CRSP at the time of the issue, and it has been listed on CRSP for at least one year (12 months) prior to the issue month; the firm is traded on NYSE, AMEX or NASDAQ.

(3) Firms must be present on COMPUSTAT (primary industrial, supplementary industrial, tertiary, full coverage and industrial research) tape for the fiscal year of offering. Accounting data necessary to compute book value, leverage must be available in COMPUSTAT. Observations with negative book value of equities are excluded.

(4) The company is not a regulated utility;

(5) The issue is a primary seasoned offering (offerings including any secondary shares are excluded. The offer must be a cash offer of common stock (joint offerings and unit offerings are excluded). Also the issue is a firm commitment, underwritten offering.

We exclude SEOs by the same firm during the five years after an SEO that is in our sample. Thus, once a firm has a seasoned equity offering the firm cannot reenter the SEO sample until five years from the issue date have passed. We collect annual data reported as of data year t-1 in COMPUSTAT, which usually become available in mid-year t, and use the latest available data for each observation. Book to market ratio, financial leverage, asset growth rate and OIBD/asset are calculated from COMPUSTAT data, and firm's size is made by production of share price and outstanding share.

For each year, some new firm might be listed on the market, and old firms became unlisted. All variables are calculated yearly. Table 3.5 presents the number of SEO firms from year 1999 to 2009. From table 3.5, we notice that 2009 has more SEO

firms relative to other years in the sample period with the least number of SEO's occurring in 2008.

Table 3.5 Number of Seasoned Equity Offering (SEO) by Calendar Year

year	NO. of SEO	Percentage of sample
1999	87	8.08%
2000	105	9.75%
2001	70	6.50%
2002	79	7.34%
2003	103	9.56%
2004	108	10.03%
2005	74	6.87%
2006	81	7.52%
2007	95	8.82%
2008	57	5.29%
2009	218	20.24%
Total	1077	100.00%

*The sample includes CRSP-listed Nasdaq, AMEX, and NYSE firms. SEOs must have at least some shares issued by the company to be included in the sample. An SEO is excluded if the issue date is within five years after an SEO by the same firm that is in our sample. Regulated utilities (SIC=481 and 491-494) are excluded.

In this scenario of seasoned equity offerings, firms choose whether to issue seasoned equity or not, according to their own perceived circumstances, thus it is a non-experimental dataset. The treatment is seasoned equity offering (SEO), the treated group includes firms that had a SEO while the control group is composed of firms that did not issue a SEO. The variable of interest is OIBD/asset. We use book to market ratio, financial leverage, asset growth rate, firm size, their squared terms and their cross terms as my coefficient drivers, because book to market ratio, financial leverage, asset growth rate, firm size are commonly believed to affect both firm's operational performance OIBD/asset, and the firm's decision to issue new equity.

3.4.2.2. Financial Validation of Method

To check our new method's validation in finance markets, we randomly assign 0 or 1 to each observation as their SEO choice rather than using their true SEO decision, we call my random assignment pseudo-SEO. If the observation is assigned 1, then we say it is treated and is in the SEO group, while if the unit gets 0 then it is not treated and in the non-SEO control group. Because the assignment of 0 or 1 is random, we should expect to see that there is no difference in OIBD/asset between the treated group and control group, which could be interpreted as the expectation of pseudo-SEO's impact on OIBD/asset is zero if this method is a valid method to estimate treatment effects in finance market. We use the mean of SEO's treatment effects to gauge the expectation of the treatment impact on OIBD/asset. We also report the positive percentage of the impact distribution for reference.

Assuming percentage of treated (SEO) units occupying 10%, 20%, … , 90% respectively of the whole sample, we repeat the above procedure 1000 times for each combination of year and treated percentage. Table 3.6A and 3.6B report the corresponding mean and positive percentage of impact distribution together with their standard error. For each cell in table 5A, we cannot reject the hypothesis that the mean is equal to zero, so we say expectation of pseudo-SEO's impact on OIBD/asset is zero. Moreover, the positive percentages reported in Table 3.6B are very close to 50%, and we cannot reject the hypothesis that this percentage is equal to 50% with statistics test. Thus we find that half of firms benefit from pseudo-SEO in terms of their OIBD/asset, while half of them do not, which would appear to validate this method in this empirical setting.

Table 3.6 Mean of pseudo-SEO's impact distribution

percentage of treated	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
10%	-0.0054 (0.0161)	-0.0062 (0.0281)	-0.0085 (0.0331)	-0.0159 (0.0320)	-0.0361 (0.0442)	-0.0223 (0.0296)	-0.0077 (0.0188)	-0.0190 (0.0330)	-0.0125 (0.0263)	-0.0610 (0.0882)	-0.0080 (0.0217)
20%	-0.0004 (0.0084)	-0.0015 (0.0073)	-0.0033 (0.0095)	-0.0145 (0.0288)	-0.0172 (0.0245)	-0.0123 (0.0193)	-0.0029 (0.0091)	-0.0075 (0.0138)	-0.0023 (0.0129)	-0.0079 (0.0240)	-0.0019 (0.0133)
30%	-0.0013 (0.0068)	-0.0015 (0.0093)	-0.0011 (0.0085)	-0.0030 (0.0137)	-0.0108 (0.0182)	-0.0069 (0.0179)	0.0001 (0.0081)	-0.0059 (0.0096)	-0.0024 (0.0096)	-0.0080 (0.0208)	-0.0006 (0.0103)
40%	-0.0010 (0.0062)	0.0014 (0.0068)	0.0019 (0.0073)	-0.0044 (0.0117)	-0.0044 (0.0143)	-0.0037 (0.0094)	-0.0021 (0.0068)	-0.0015 (0.0089)	0.0000 (0.0086)	-0.0031 (0.0142)	-0.0012 (0.0099)
50%	-0.0002 (0.0062)	-0.0001 (0.0053)	0.0002 (0.0075)	-0.0013 (0.0124)	-0.0001 (0.0155)	0.0017 (0.0093)	-0.0007 (0.0053)	-0.0004 (0.0100)	-0.0007 (0.0071)	0.0012 (0.0112)	0.0002 (0.0095)
60%	0.0015 (0.0070)	0.0012 (0.0066)	-0.0007 (0.0083)	0.0018 (0.0098)	0.0031 (0.0166)	0.0016 (0.0072)	-0.0008 (0.0064)	0.0027 (0.0088)	0.0005 (0.0102)	0.0025 (0.0109)	0.0015 (0.0095)
70%	0.0021 (0.0069)	0.0023 (0.0096)	0.0000 (0.0076)	0.0056 (0.0123)	0.0095 (0.0192)	0.0035 (0.0117)	0.0031 (0.0086)	0.0059 (0.0128)	0.0005 (0.0093)	0.0088 (0.0250)	0.0008 (0.0115)
80%	0.0020 (0.0079)	0.0039 (0.0115)	0.0010 (0.0091)	0.0101 (0.0233)	0.0167 (0.0221)	0.0084 (0.0167)	0.0042 (0.0107)	0.0085 (0.0136)	0.0042 (0.0152)	0.0195 (0.0352)	0.0007 (0.0118)
90%	0.0043 (0.0111)	0.0079 (0.0215)	0.0040 (0.0186)	0.0128 (0.0434)	0.0367 (0.0420)	0.0251 (0.0357)	0.0054 (0.0151)	0.0238 (0.0297)	0.0142 (0.0247)	0.0688 (0.1133)	0.0092 (0.0210)

* Numbers in parenthesis are standard errors.

Table 3.7 Positive percentage of pseudo-SEO's impact distribution

percentage of treated	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
10%	0.5213 (0.1249)	0.5282 (0.1161)	0.5184 (0.1213)	0.5237 (0.0975)	0.5740 (0.1016)	0.5074 (0.0933)	0.5315 (0.1057)	0.5276 (0.1056)	0.5277 (0.0882)	0.5372 (0.0851)	0.5129 (0.0920)
20%	0.5302 (0.1039)	0.5084 (0.1075)	0.4855 (0.1070)	0.5317 (0.1173)	0.5708 (0.0721)	0.5040 (0.1104)	0.5120 (0.1125)	0.5243 (0.1051)	0.5117 (0.0941)	0.5256 (0.1013)	0.5321 (0.0819)
30%	0.5140 (0.1078)	0.4935 (0.1020)	0.5017 (0.1276)	0.4893 (0.1229)	0.5396 (0.1030)	0.5250 (0.0876)	0.5458 (0.0960)	0.4942 (0.0935)	0.4956 (0.1058)	0.5041 (0.1020)	0.5147 (0.0935)
40%	0.5079 (0.1017)	0.5218 (0.1012)	0.5279 (0.1171)	0.4913 (0.0898)	0.5429 (0.1003)	0.4909 (0.1038)	0.4479 (0.1022)	0.5060 (0.1000)	0.5001 (0.1077)	0.4889 (0.1051)	0.5112 (0.0926)
50%	0.5003 (0.0999)	0.4985 (0.1071)	0.5145 (0.1200)	0.4796 (0.0989)	0.5284 (0.1018)	0.5390 (0.1035)	0.4919 (0.1039)	0.5024 (0.1011)	0.4961 (0.0964)	0.5215 (0.1136)	0.5014 (0.0996)
60%	0.4956 (0.1084)	0.5229 (0.1116)	0.4899 (0.1446)	0.5095 (0.1018)	0.5017 (0.1052)	0.5134 (0.1164)	0.4753 (0.1213)	0.4950 (0.0998)	0.4959 (0.1109)	0.4963 (0.1097)	0.4896 (0.0965)
70%	0.5044 (0.1120)	0.5199 (0.1135)	0.4876 (0.1233)	0.4940 (0.1224)	0.4897 (0.0938)	0.4760 (0.0939)	0.4973 (0.1140)	0.4921 (0.1139)	0.4881 (0.1029)	0.4885 (0.1072)	0.5059 (0.1073)
80%	0.5111 (0.1016)	0.5050 (0.1075)	0.4850 (0.1121)	0.4740 (0.0976)	0.4476 (0.0885)	0.4671 (0.0952)	0.4994 (0.1199)	0.4795 (0.1088)	0.4833 (0.0925)	0.5027 (0.0962)	0.4762 (0.0986)
90%	0.4560 (0.1042)	0.4645 (0.0827)	0.4706 (0.1205)	0.4715 (0.0939)	0.4387 (0.0921)	0.4565 (0.0998)	0.4750 (0.0965)	0.4768 (0.0971)	0.5189 (0.0999)	0.4748 (0.0996)	0.4589 (0.0912)

* Numbers in parenthesis are standard errors

3.4.2.3. Estimation With Real SEO Data

We next examine results based on the actual SEO data. For the SEO firms in each year, we check their 1 year, 3 year and 5 year post-issue performance. The model gives good R^2 s for 1 year post-issue OIBD/asset, but for 3 year and 5 year post-issue performance, the R^2 s only attain 45% at most. We thus focus only on results for 1 year post issue OIBD/asset. The quantile values for the SEO's treatment effect distribution, mean, positive percentage and R^2 for year 1999-2008 are listed in Table 6. For most cases, the impact distribution of SEO on OIBD/asset has a negative mean, negative

median and less than 50% positive percentage. For a few cases the mean of distribution exceeds zero, but the value is quite small (~< 0.005), and the corresponding positive percentage is smaller than 50%.

Table 3.8 Properties of SEO impact distribution (on 1 year post-issue OIBD/asset)

Year	5^{th} percentile	25^{th} percentile	50^{th} percentile	7 5^{th} percentile	95^{th} percentile	Mean	percent positive	R^2
1999	-0.4401	-0.1884	-0.0563	0.0542	0.2804	-0.0615	0.3587	0.6316
2000	-0.1111	-0.0566	-0.0163	0.0228	0.1660	-0.0162	0.3910	0.6082
2001	-0.2419	-0.1112	-0.0409	0.0080	0.0942	-0.0589	0.2869	0.6360
2002	-0.1811	-0.0406	-0.0074	0.0225	0.1360	-0.0303	0.4231	0.6481
2003	-0.2298	-0.0762	-0.0221	0.0031	0.0464	-0.0511	0.2892	0.7276
2004	-0.1247	-0.0438	-0.0162	0.0039	0.0474	-0.0100	0.2958	0.6264
2005	-0.0494	-0.0218	0.0060	0.0462	0.1624	0.0035	0.4537	0.7412
2006	-0.1840	-0.0488	-0.0046	0.0855	0.3463	0.0051	0.4754	0.7165
2007	-0.1604	-0.0482	-0.0137	0.0196	0.1116	-0.0159	0.3820	0.7282
2008	-0.3030	-0.0990	-0.0227	0.0490	0.2391	-0.0052	0.4096	0.5687

Smoothed densities of the treatment effect distributions are plotted in Figure 3.12-3.13 which shows quite clearly that the impact distribution has substantial negative support and very little positive support, suggesting that SEO firms have underperformed based on the OIBD/asset measure.

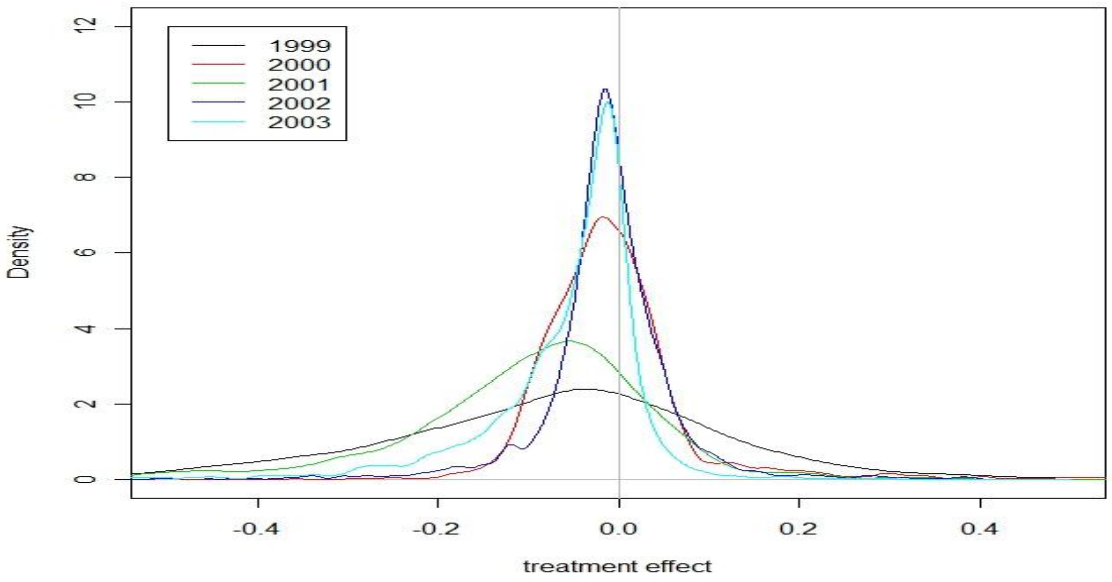

Figure 3.12 Impact distribution of SEO on 1 year post-issue OIBD/asset

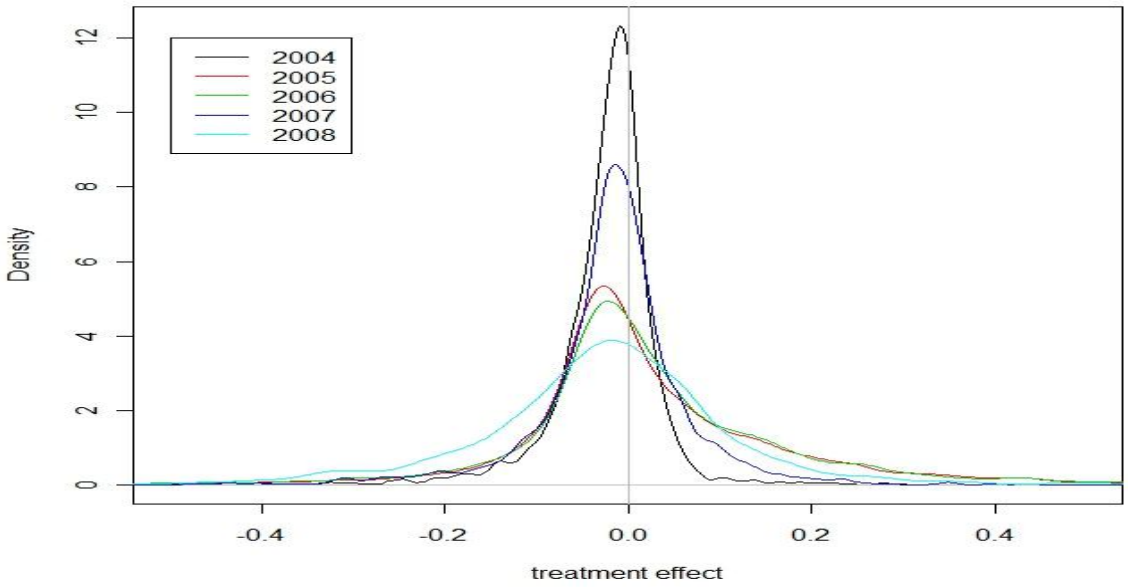

Figure 3.13 Impact distribution of SEO on 1 year post-issue OIBD/asset (continued)

3.5. Conclusions

When observations are self-selected into a treatment according to their characteristics, treatment assignment is not completely random. In this case the average treatment effect and the average treatment effect on the treated are hard to compute as the computation involves an unobservable counterfactual. Furthermore, the estimations of individual treatment effect and treatment effect distribution cannot be constructed. This paper suggests a method in which either the endogeneity of treatment assignment or the unobservable counterfactual can be addressed. This claim is supported for large samples on theoretical grounds and as well as by finite results of simulation experiments wherein each unit is self-selected into treatment according to an index function depending on unit's own characteristics. The CVC estimator tracks individual treatment effects closely and also the impact distribution. This method is also applied in two empirical settings. With the JTPA study dataset, the estimator states more than 60% people benefit from training program, which is the same as Heckman (1997) estimated with a similar dataset. In the financial setting, examining the performance of firms that had seasoned equity offerings, the method points to the poor relative performance of SEO firms' post-issue OIBD/asset.

References

[1]. Abadie, A., Angrist, J. and Imbens G. (2002), Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings, Econometrica, 70, 91–117.

[2]. Athey, S. and Imberns G. (2006), Identification and Inference in Nonlinear Difference-in-Difference Models, Econometrica, 74, 431-497.

[3]. Abadie, A. and Imbens G. (2006), Large Sample Properties of Matching Estimators for Average Treatment Effects, Econometrica, 74, 235–267.

[4]. Basmann, R.L. (1988), Causality Tests and Observationally Equivalent Representations of Econometric Models, Journal of Econometrics, Annals, 39, 69-104.

[5]. Becker, S.O. and Ichino, A. (2002), Estimation of Average Treatment Effects Based on Propensity Socre, Stata Journal 2(4), 358-377.

[6]. Becker, S.O. and Caliendo, M. (2007). Sensitivity Analysis for Average Treatment Effect, Stat Journal 7(1), 71-83.

[7]. Bontemps, C., Racine, J.S. and Simioni, M. (2009), Nonparametric vs Parametric Binary Choice Models: An Empirical Investigation, 2009 Annual Meeting, July 26-28, 2009, Milwaukee, Wisconsin 49286, Agricultural and Applied Economics Association.

[8]. Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J.,and Stürmer, T. (2006), Variable Selection for Propensity Score Models, American Journal of Epidemiology, 163, 1149–1156.

[9]. Bryson, A., Dorsett, R. and Purdon, S. (2002), The Use of Propensity Score Matching in the Evaluation of Labour Market Policies, Working Paper No. 4, Department for Work and Pensions.

[10]. Buchmueller, T., Grumbach, K., Kronick R. and Kahn, J. (2005), The Effect of Health Insurance on Medical Care Utilization and Implication for Insurance Expansion: A Review of the Literature, Medical Care Research and Review, 62(1), 3-30.

[11]. Caliendo, M., and Kopeinig S. (2008), Some Practical Guidance for the Implementation of Propensity Score Matching, Journal of Economic Surveys, 22(1), 31-72.

[12]. Card, D., Dobkin, C. and Maestas N. (2008), The Impact of Nearly Universal Insurance Coverage on Health Care Utilization and Health: Evidence from Medicare, American Economic Review, 98(5), 2242-2258.

[13]. DeNavas-Walt, C., Bernadette B.D. Proctor, and Smith J.C. (2009), Income, Poverty, and Health Insurance Coverage in the United States, 2008, U.S. Census Bureau Report, 60-236.

[14]. Chang, I., Hallahan, C. and Swamy, P.A.V.B. (1992), Efficient Computation of Stochastic Coefficients Models, Computational Economics and Econometrics, 43-53.

[15]. Chang, I., Swamy, P.A.V.B., Hallahan, C. and Tavlas, G. S. (2000), A Computational Approach to Finding Causal Economic Laws, Computational Economics, 16, 105-136.

[16]. Chernozhukov, V. and Hansen C. (2005), An IV Model of Quantile Treatment Effects, Econometrica, 73, 245-261.

[17]. Chernozhukov, V. and Hansen C. (2006), Instrumental Quantile Regression Inference for Structural and Treatment Effect Models, Journal of Econometrics, 132, 491–525.

[18]. Chung, C. and Goldberger A.S. (1984), Proportional Projections in Limited Dependent Variable Models, Econometrica, 52, 531-534.

[19]. Currie, J. and Gruber J. (1996), Health Insurance Eligibility, Utilization of Medical Care, and Child Health, the Quarterly Journal of Economics, 111(2), 431-466.

[20]. De Finetti, B. (1974), The Theory of Probability, Vol.1, New York: John Wiley & Sons.

[21]. Dehejia, R.H. and Wahba, S. (1999), Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs, Journal of the American Statistical Association 94 (11), 1053-1062.

[22]. Dehejia, R.H. and Wahba, S. (2002), Propensity Score Matching Methods for Nonexperimental Causal Studies, Review of Economics and Statistics, 84(1), 151-161.

[23]. Dehejia, R.H. (2005), Practical Propensity Score Matching: a Reply to Smith and Todd, Journal of Econometrics, 125(1-2), 355-364.

[24]. DeNavas-Walt, C., Bernadette B.D. Proctor, and Smith J.C (2010), Income, Poverty, and Health Insurance Coverage in the United States, 2009, U.S. Census Bureau Report, 60-238.

[25]. Fan, J. (1992a), Design Adaptive Nonparametric Regression, Journal of the American Statistical Association, 87, 998-1004.

[26]. Fan, J. (1992b), Local linear Regression Smoothers and their Minimax Efficiencies, The Annals of Statistics, 21, 196-216

[27]. Fan, Y. and Park S. (2009), Partial Identification of the Distribution of Treatment Effects and its Confidence Sets, Advances in Econometrics, 25, 3–70.

[28]. Fan, Y. and Wu J. (2010): Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and its Confidence Sets, Review of Economic Studies , 77(3), 1002— 1041.

[29]. Finkelstein, A. (2007), The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare, The Quarterly Journal of Economics, 122(1), 1-37

[30]. Firpo, S. (2007), Effcient Semiparametric Estimation of Quantile Treatment Effects, Econometrica, 75, 259-276

[31]. French, E. and Kamboj, K. (2002), Analyzing the Relationship between Health Insurance, Health Costs, and Health Care Utilization, Economic Perspectives, 26, 60-72.

[32]. Greene, W. H. (2008): Econometric Analysis (Pearson Prentice Hall, Upper Saddle River, New Jersey).

[33]. Guo, R., Sickles, R.C. and Swamy, P.A.V.B. (2011), The Use of Cross-Sectionally Varying Coefficient Models in the Estimation of Treatment Effects, working paper.

[34]. Hahn, J. (1998), On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects, Econometrica, 66, 315–331.

[35]. Havenman, R. and Wolfe, B.(2010), US Health Care Reform: A Primer and An Assessment, CESifo DICE Report, Vol.8, Issue 3, 53-60.

[36]. Heckman, J. (1979), Sample Selection Bias as a Specification Error, Econometrica, 47(1), 153-161.

[37]. Heckman, J., Ichimura, H. and Todd, P. (1997), Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program, Review of Economic Studies, 64, 605–654.

[38]. Heckman, J., Ichimura, H. and Todd, P. (1998), Matching as an Econometric Evaluation Estimator, Review of Economic Studies, 65, 261–294.

[39]. Heckman, J., Smith, J. and Clements, N. (1997b), Making the Most Out of Social Experiments: Accounting for Heterogeneity in Programme Impacts, Review of Economic Studies 64, 487–536.

[40]. Heckman, J. and Vytlacil, E. (2005), Structural Equations, Treatment Effects and Econometric Policy Evaluation, Econometrica, 73, 669-738.

[41]. Heinrich, C.J., Meuser, P.R. and Troske, K.R. (2005), Welfare to Temporary Work: Implications for Labor Market Outcomes, The Review of Economics and Statistics, 87(1), 154-173.

[42]. Hirano, K., Imbens, G. and Ridder, G. (2003), Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, Econometrica, 71, 1161–1189.

[43]. Holland, P.W. (1986), Statistics and Causal Inference, Journal of the American Statistical Association, 81,945-970.

[44]. Holland, P.W. (1988), Causal Inference in Retrospective Studies, Evaluation Review, 12(3), 203-231.

[45]. Huber, M., M. Lechner, and C. Wunsch (2010): How to control for many covariates? Reliable estimators based on the propensity score, Discussion paper 2010-30, Department of Economics, University of St. Gallen.

[46]. Imbens, G. (2004), Nonparametric Estimation of Average Treatment Effects under Exogeneity: A review, Review of Economics and Statistics, 86, 4-29.

[47]. Imbens, G. and Wooldridge, J. (2009), Recent Developments in the Econometrics of Program Evaluation, Journal of Economic Literature, 47(1), 5-86.

[48]. Jalan, J. and Ravallion, M. (2003), Estimating the Benefit Incidence of an Antipoverty Program by Propensity Score Matching, Journal of Business & Economics Statistics, 21(1), 19-30.

[49]. Klein, R.W. and Spady, R.H. (1993), An Efficient Semiparametric Estimator for Binary Response Models, Econometrica, 61, 387-421.

[50]. Kolstad, J. and Kowalski, A. (2010), The Impact of Heal Care Reform on Hospital and Preventive Care: Evidence from Massachusetts, National Bureau of Economic Research Working Paper #16012.

[51]. Kordas, G. and Lehrer, S.F. (2004), Matching using Semiparametric Propensity Scores, No 441, Econometric Society 2004 North American Summer Meetings

[52]. LaPlante, M.P. (1993), Disability, Health Insurance Coverage, and Utilization of Acute Health Services in the United States, Disability Statistics Report No.4. Washington, D.C. U.S. Department of Education National Institute on Disability and Rehabilitation Research.

[53]. Lee, W. (2006), Propensity Score Matching and Variations on the Balancing Test, Technical report, Melbourne Institute of Applied Economic and Social Research.

[54]. Lehmann, E. and Casella, G. (1998), Theory of Point Estimation, Second edition, New York: Springer.

[55]. Levin, D. and Painter, G. (2003), The Schooling Costs of Teenage Out-of-Wedlock Childbearing: Analysis with a Within-School Propensity-Score-Matching Estimator, The Review of Economics and Statistics, 85(4), 884-900.

[56]. Li, Q., Racine, J. and Wooldridge, J. (2009), Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data, Journal of Business & Economic Statistics, 27(2), 206-223.

[57]. Lillie-Blanton, M. (2008), Addressing Disparities in Health and Healthcare: Issues for Reform, Testimony before Congress House of Representatives Committee on Ways and Means Health SubCommittee June 10.

[58]. Loughran, T. and Ritter, J.R. (1995), The New Issues Puzzle, Journal of Finance, 50, 23-51.

[59]. Loughran, T. and Ritter, J.R. (1997): The Operating Performance of Firms Conducting Seasoned Equity Offerings, Journal of Finance, 52, 1823-1850.

[60]. Loughran, T. and Ritter, J.R. (2000), Uniformly Least Powerful Tests of Market Efficiency, Journal of Financial Economics, 55, 361-389.

[61]. Millimet, D.L. and Tchernis, R. (2009), On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies, Journal of Business & Economic Statistics, 27, 297-315.

[62]. Neyman, J. (1923), Sur les Applications de la thar des Probabilities aux Experiences Aggaricales: Essay des Principle. Excerpts reprinted (1990) in English (D. Dabrowska and T. Speed, Trans.) in Statistical Science 5, 463-472.

[63]. Pratt, J. and Schlaifer, R. (1984), On the Nature and Discovery of Structure, Journal of the American Statistical Association, 79, 9-21, 29-33.

[64]. Pratt, J. and Schlaifer, R. (1988), On the Interpretation and Observation of Laws, Journal of Econometrics, 39, 23-52.

[65]. Robins, J. and Rotnitzky, A. (1995), Semiparametric Efficiency in Multivariate Regression Models with Missing Data, Journal of the American Statistical Association, 90 , 122–129.

[66]. Rubin, D. (1974), Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, Journal of Educational Psychology, 66, 688-701.

[67]. Rubin, D. (1977), Assignment to a Treatment Group on the Basis of a Covariate, Journal of Educational Statistics, 2, 1-26.

[68]. Rubin, D. (1980), Bias Reduction Using Mahalanobis's Metric Matching, Biometrics, 36, 295-298.

[69]. Rubin, D. and Thomas, N. (1996), Matching Using Estimated Propensity Scores: Relating Theory to Practice, Biometrics, 52, 249-264.

[70]. Rosenbaum, P. (1995) Observational Studies (New York: Springer-Verlag).

[71]. Rosenbaum, P. and Rubin, D. (1983), The Central Role of Propensity Score in Observational Studies for Causal Effects, Biometrika, 70, 41-55.

[72]. Sekhon, J.S. (2011), Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. Journal of Statistical Software, 42(7), 1-52.

[73]. Sickles, R.C., and Taubman, P. (1986), An Analysis of the Health and Retirement Status of the Elderly, Econometrica, 54, 1339-1356.

[74]. Skyrms, B. (1988), Probability and Causation, Journal of Econometrics, 39, 53-68.

[75]. Smith, J. and Todd, P. (2005), Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators? Journal of Econometrics, 125, 305-353.

[76]. Swamy, P.A.V.B. and Mehta, J. (1975), Bayesian and non-Bayesian Analysis of Switching Regressions and of Random Coefficient Regression Models, Journal of the American Statistical Association 70, 593-602.

[77]. Swamy, P.A.V.B. and Mehta, J. (1976), Minimum Average Risk Estimators for Coefficients in Linear Models, Communications in Statistics, 5(9), 803-818.

[78]. Swamy, P.A.V.B. and Mehta, J. (1977a), Estimation of Linear Models with Time and Cross-Sectionally-Varying Parameters, Journal of the American Statistical Association, 72, Dec., 890-891.

[79]. Swamy, P.A.V.B. and Tinsley, P. (1980), Linear Prediction and Estimation Methods for Regression Models with Stationary Stochastic Coefficients, Journal of Econometrics, 12, 103-142.

[80]. Swamy, P.A.V.B., Tavlas, G.S., Hall, S.G. and Hondroyiannis, S. (2010), Estimation of Parameters in the Presence of Model Misspecification and Measurement Error, Studies in Nonlinear Dynamics & Econometrics, 14, 1-33.

[81]. Swamy, P. A. V. B. and Hall, S.G. (2011), Measurement of Causal Effects, Economic Change and Restructuring, 45, 3-23 forthcoming

[82]. Wu, X. and Perloff, S. (2006), Information-Theoretic Deconvolution Approximation of Treatment Effect Distribution, typescript.

[83]. Zhao, Z. (2008), Sensitivity of Propensity Score Methods to the Specifications, Economics Letters, 98, 309–319